José Salinas Ortiz

Análisis estadísticos para la toma de decisiones en administración y economía

Biblioteca Universitaria

Los economistas y administradores en el sector privado y público se enfrentan constantemente a decisiones. Para tomar estas decisiones, ellos deben definir las alternativas y deben buscar la información relevante y estructurarla. A menudo, la información es numérica y el decisor debe pensar en una manera de condensarla y estudiarla. Algunas veces, la información es insuficiente y el decisor debe considerar si la información disponible puede ser usada para darse una idea de los hechos desconocidos.

El análisis estadístico, tema de este libro, nos enseña cómo condensar y analizar la información y cómo utilizar información incompleta para realizar pruebas y hacer predicciones acerca de los hechos que todavía no conocemos. En este libro se presentan los conceptos y técnicas estadísticas en una forma sistemática para que el lector tenga un medio de lograr un entendimiento sólido del rol del análisis estadístico en el proceso de toma de decisiones, tanto a nivel empresarial como de gobierno.

Biblioteca Universitaria / Instrumentos matemáticos para la toma de decisiones / 12

José Salinas Ortiz

Análisis estadístico para la toma de decisiones en administración y economía

UNIVERSIDAD DEL PACIFICO

BUP-CENDI

Salinas Ortiz, José
Análisis estadístico para la toma de decisiones en administración y economía. -- Lima : Universidad del Pacífico, 1993.

/ANÁLISIS ESTADÍSTICO/TOMA DE DECISIONES/ADMINISTRACIÓN/ECONOMÍA/

519.24:65 (CDU)

Avenida Salaverry 2020, Lima 11, Perú
Primera edición, diciembre de 1993

Miembro de la Asociación Peruana de Editoriales Universitarias y de Escuelas Superiores (APESU), y de la Asociación de Editoriales Universitarias de América Latina y el Caribe (EULAC).

La publicación de la Biblioteca Universitaria se realiza gracias al Proyecto de Mejoramiento Gerencial del Sector Privado, financiado por la Agencia para el Desarrollo Internacional del Gobierno de los Estados Unidos de Norteamérica y administrado por la Asociación Perú Texas.

Diseño gráfico: Carlos Tovar Samanez
Cuidado de la edición: José Luis Carrillo Mendoza
Impreso en el Perú - Printed in Peru
I.S.B.N. 84-89293-75-9

Contenido

Presentación de la Biblioteca Universitaria

Nuestra institución cree firmemente que las organizaciones tendrán éxito en la medida en que satisfagan necesidades existentes en una sociedad. Así, en un estudio analizamos el comportamiento de alumnos y profesores ante el mercado de textos universitarios, y observamos que la mayoría de profesores recomiendan textos extranjeros debido a la carencia de material bibliográfico peruano referido a nuestra realidad. Además, manifestaron requerir libros que se encuentren metodológicamente bien organizados. Esta necesidad nos motivó a iniciar la elaboración y publicación de veintiséis textos para educar a universitarios peruanos. Las ventajas competitivas de los mismos serán justamente aquellas requeridas por el mercado: estar referidos a nuestra realidad y presentarse de manera que faciliten el aprendizaje.

Las obras versarán sobre temas relacionados a la agroempresa, negocios internacionales, contabilidad y finanzas, mercadotecnia, recursos humanos, análisis económico para la empresa, sistemas para la toma de decisiones, entre otros.

La realización de esta actividad ha sido posible gracias a la participación de un equipo de profesionales y al Proyecto de Mejoramiento Gerencial del Sector Privado que se desarrolla con el apoyo de la Agencia para el Desarrollo Internacional del Gobierno de los Estados Unidos de Norteamérica (AID).

Lima, febrero de 1992

Estuardo Marrou Loayza
Coordinador del Proyecto de Mejoramiento
Gerencial del Sector Privado
Universidad del Pacífico

Prólogo

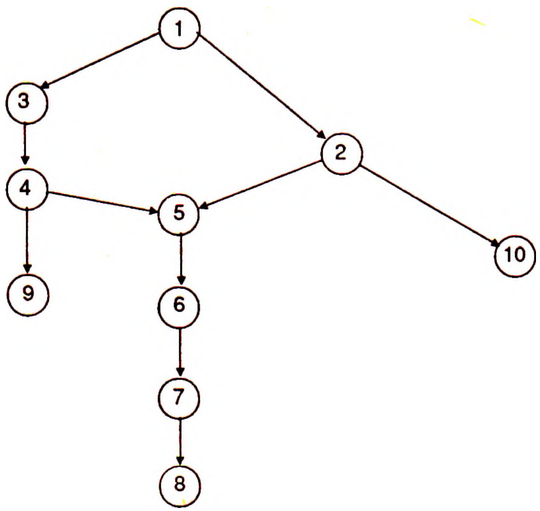
Este libro es el resultado del desarrollo de las notas de clase usadas por el autor en el dictado del curso Métodos Cuantitativos I en la Escuela de Postgrado de la Universidad del Pacífico durante los últimos cinco años. Está escrito como texto para un curso de Análisis Estadístico, para estudiantes de postgrado del primer nivel o para no graduados de segundo nivel en las áreas de administración y economía.

El objetivo de este texto es familiarizar a los estudiantes de administración y economía con los conceptos y técnicas estadísticas y su aplicación al mundo de hoy, en una forma accesible y sin prerrequisitos matemáticos rigurosos. De esta manera se espera brindar a los estudiantes un medio para lograr un entendimiento conceptual sólido del papel de las técnicas estadísticas en el proceso de toma de decisiones. Se mostrará cómo estas técnicas ayudan tanto a los economistas como administradores a tomar decisiones en los sectores privado y público a través del desarrollo de su habilidad para analizar diferentes factores que influyen en el resultado de todo problema de decisión.

Se describirán las técnicas estadísticas, su operatividad y aplicación directa por el decisor. El enfoque es eminentemente práctico. Cada capítulo presenta, en primer lugar, una situación-problema que busca motivar al lector a utilizar su ingenio y creatividad en el planteamiento de soluciones innovativas. A continuación se describe la técnica y modelo cuantitativo.

El desarrollo de la técnica cuantitativa o modelo incluye su aplicación a la situación-problema para poder generar una solución o decisión. Se espera que este enfoque motive al estudiante al demostrar no sólo cómo puede ser aplicado el procedimiento, sino también cómo contribuye al proceso de toma de decisiones.

El libro contiene material suficiente para un curso de un ciclo académico. Sin embargo, hay flexibilidad para desarrollar un plan de estudio. En la Escuela de Postgrado de la Universidad del Pacífico el libro ha servido para un ciclo completo. El siguiente esquema de la dependencia de los capítulos puede usarse para diseñar un programa específico.



Un componente importante del libro es el conjunto de problemas que aparece al final de cada capítulo. Los problemas son ejercicios, esto es, una aplicación más o menos directa de las técnicas discutidas en el capítulo.

La preparación de este libro ha sido una tarea muy larga, imposible de terminar sin la ayuda de mucha gente. Me gustaría agradecer a la Escuela de Postgrado de la Universidad del Pacífico, que brindó el ambiente y recursos para hacer este proyecto posible. También quiero agradecer a todos los alumnos que con paciencia supieron entender y ayudar a superar las dificultades asociadas con la evolución de las notas de clase, y que desinteresadamente contribuyeron a superarlas.

José Salinas Ortiz

Introducción

1. El problema de la toma de decisiones. 2. La incertidumbre y el análisis estadístico. 3. El análisis estadístico y la información disponible. 4. Relación entre variables y predicción. 5. La teoría de decisiones.

1. EL PROCESO DE TOMA DE DECISIONES

El medio ambiente en el que se toman las decisiones humanas, tanto individuales como organizacionales, se caracteriza por ser incierto, complejo, dinámico, competitivo y finito. Este entorno es *incierto* porque el decisor no puede estar seguro del comportamiento futuro de algunos factores que influyen en el resultado de la decisión. Estamos obligados a enfrentar un mundo *complejo* y *dinámico* donde existen muchos factores que interactúan en formas que no son fácilmente comprensibles y que evolucionan continuamente a través del tiempo. Finalmente, debido a que la disponibilidad de recursos es limitada, el entorno es *finito*, lo que trae como consecuencia un alto grado de *competitividad* entre los agentes económicos, pues cada quien tratará de obtener el mayor beneficio para sí. Todas estas características producen, con frecuencia, un sentimiento de confusión y preocupación en el decisor.

Para hacer frente a las características de su entorno, el hombre dispone de ciertos recursos que pueden facilitarle tomar decisiones que le produzcan el bienestar que desea. En primer lugar, puede utilizar su *ingenio* para concebir y formular diferentes cursos de acción, definiendo sus alternativas potenciales. Adicionalmente, a través de su *percepción* puede aprender de lo que experimenta y acumular información del medio ambiente que lo rodea. Finalmente, dispone de una *filosofía*, un cúmulo de principios que guían su vida y le permiten establecer sus preferencias con respecto a los varios resultados que podría generar una decisión.

En muchos casos, tanto en decisiones triviales como en aquellas que involucran a una corporación, o aun a toda una nación, el decisor se vale de un proceso *intuitivo* para considerar sus alternativas de acción a la luz de la información acumulada y sobre la base de sus preferencias. La lógica de las decisiones tomadas intuitivamente es incomprobable; en algunos casos, el decisor no será capaz de explicar las razones por las cuales tomó una decisión en particular.

En la sociedad moderna, donde se verifica una mayor interdependencia entre los diferentes actores económicos –familias, empresas y gobiernos–, las decisiones importantes pueden y deben tomarse utilizando herramientas que vayan más allá de la intuición; que sean transparentes, explícitas y que tengan una lógica comprobable. Los diferentes métodos cuantitativos del análisis estadístico ayudan al decisor al permitirle explicitar y describir sus *alternativas* cuantitativamente, sistematizar y estructurar su *información* disponible, y considerar explícitamente sus *preferencias*.

El rápido progreso de la tecnología de las computadoras, combinado con el desarrollo de procedimientos computarizados refinados en los últimos años, ha ampliado el rango de problemas complejos que pueden ser tratados confiable y eficientemente por los diferentes métodos cuantitativos. Por otro

lado, las crisis recientes en áreas como energía, productividad industrial, balanza de pagos y en la economía en general, nos obligan a evaluar más cuidadosamente las diferentes alternativas de decisión considerando toda la información relevante. La necesidad de una evaluación meticulosa ha producido nuevos intereses en conocer y utilizar herramientas metodológicas que sean capaces de generar decisiones óptimas en problemas importantes y cruciales para nuestra existencia.

En este libro se revisarán los conceptos y técnicas del análisis estadístico y su aplicación al proceso de toma de decisiones en una forma accesible a estudiantes que no tienen una formación matemática rigurosa.

2. LA INCERTIDUMBRE Y EL ANÁLISIS ESTADÍSTICO

Una característica importante de nuestro medio ambiente es la incertidumbre que rodea al comportamiento de los elementos que contribuyen al resultado de una decisión. La incorporación de los factores inciertos en el análisis de decisiones es posible gracias a la teoría de probabilidades, que proporciona el razonamiento cuantitativo para entender fenómenos aleatorios. Además de permitir el análisis de incertidumbre en forma explícita, el análisis probabilístico también es la base de la inferencia estadística.

Muchos estudiantes han tenido dificultades en el estudio de la teoría de probabilidades, limitándolo al aprendizaje y aplicación de fórmulas y reglas. El enfoque usado en este libro pretende explicar la teoría de probabilidades apelando a la intuición gráfica del lector sin incluir conceptos matemáticos intrincados. El desarrollo de esta teoría se presenta utilizando el lenguaje especial del álgebra de eventos, que permite describir en forma sencilla y clara los eventos bajo consideración. El álgebra de eventos facilita establecer relaciones que definen el concepto de probabilidad, probabilidad condicional e independencia. Final-

mente, se introduce el concepto de variable aleatoria utilizando el gráfico del árbol de probabilidades y se describen las variables aleatorias más importantes en el análisis estadístico de problemas en las áreas de administración y economía: las distribuciones binomial, de Poisson y normal. Esta última es la base de la estadística inferencial y del análisis de regresión.

La teoría de probabilidades se presenta en el capítulo II, y en el capítulo III se describen las variables aleatorias relevantes.

3. EL ANÁLISIS ESTADÍSTICO Y LA INFORMACIÓN DISPONIBLE

Con el objeto de tomar las mejores decisiones, el decisor debe contar con toda la información relevante que le sea posible conseguir. Mucha de esta información es numérica; otra será información cualitativa que podrá ser cuantificada de diversas maneras. La utilidad de esta vasta información numérica tan desagregada es virtualmente nula si se mantiene en su estado bruto. El decisor debe buscar una manera adecuada para refinarla, condensarla y presentarla de modo que sea fácilmente comprensible.

La estadística descriptiva, a través de sus técnicas gráficas y numéricas, nos enseña a presentar, resumir y analizar información numérica con el propósito de extraer de los datos algunas propiedades que describen adecuadamente la estructura del proceso y el comportamiento del sistema bajo estudio. Consideremos el caso de una empresa que está interesada en conocer el movimiento en el precio de las acciones, con el fin de establecer su política de inversiones. El listado diario de todos los precios de las acciones en el mercado bursátil le será de muy poca utilidad. Muy posiblemente, el Gerente Financiero extraerá la esencia de estos datos y los presentará a la Alta Gerencia de modo tal que las características más importantes de la información numérica sean identificadas de una manera clara y precisa.

En el capítulo I se presentan diferentes técnicas de la estadística descriptiva.

Por otro lado, existen instancias en las que la información disponible es insuficiente y el decisor debe determinar si es conveniente utilizar esa información incompleta para inferir acerca de los hechos desconocidos. Así por ejemplo, si una empresa manufacturera está considerando lanzar un nuevo producto al mercado, estará interesada en conocer el nivel probable de demanda que enfrentará su producto. Con el fin de evaluar la intención de compra de los consumidores potenciales del nuevo producto, se realizará un estudio de mercado mediante el cual se entrevistará a una pequeña parte –o muestra– de esta población. La información obtenida de las personas encuestadas servirá para estimar características de la población de la cual fue extraída.

Las técnicas de la estadística inferencial, con la ayuda de la teoría de probabilidades y el conocimiento sobre el manejo de información numérica, nos brindan las herramientas que permiten utilizar información muestral para hacer estimaciones y derivar conclusiones con respecto a la población total. Los principios básicos de la teoría de muestreo, estimación y los métodos de la prueba de hipótesis en un análisis univariable se presentan en los capítulos IV y V.

4. RELACIÓN ENTRE VARIABLES Y PREDICCIÓN

La información que el decisor acumula a través de su percepción también le permite formular algunas hipótesis respecto a la existencia de posibles relaciones entre dos o más variables que influyen en el resultado de sus decisiones. Las técnicas del análisis de regresión permiten medir el grado y naturaleza de la relación entre dichas variables, estimando los parámetros que la definen. También es posible probar hipótesis para establecer si la relación es estadísticamente significativa. Finalmente, el aná-

lisis de regresión nos permite predecir el valor de una variable sobre la base del valor de otras variables que la explican.

Supongamos que el Gerente de Comercialización de una empresa fabricante de bebidas gaseosas está estudiando lanzar una campaña para incrementar considerablemente sus ventas, a través de una reducción de sus precios al consumidor. La teoría económica establece que, si se mantiene el resto de factores constantes, una disminución en el precio de un producto ocasionará un incremento en su demanda. Pero la empresa necesita saber en cuánto reducir su precio para lograr un incremento de demanda tal que incremente sus utilidades totales. Para estimar esta relación hipotética entre precio y cantidad debemos contar con información numérica que demuestre cómo el volumen de ventas ha reaccionado ante cambios en el precio en el pasado. Una vez estimada la relación podremos predecir el comportamiento de la demanda ante variaciones en el precio, bajo la premisa de que lo que ocurrió en el pasado se repetirá en el futuro.

Además del análisis de regresión, existen otras técnicas de predicción para proyectar el comportamiento futuro de variables de interés para el decisor. Estas técnicas pueden agruparse en tres grandes categorías: métodos de series de tiempo, métodos causales y métodos cualitativos. Los métodos de series de tiempo se basan en el análisis de los datos históricos de la variable que intentamos predecir con el fin de identificar sus patrones de comportamiento. Los métodos causales relacionan la variable que tratamos de predecir con otras variables que la explican. El análisis de regresión es un caso especial de los métodos causales. Finalmente, los métodos cualitativos de predicción se basan primordialmente en información cualitativa, como puede ser la opinión de expertos en el tema de interés.

El análisis de regresión se presenta en el capítulo VI, y las técnicas de predicción en el capítulo VII.

5. LA TEORÍA DE DECISIONES

La característica del entorno que más problemas ocasiona al decisor es la incertidumbre. La salida más fácil es sin duda hacer uso de modelos de determinación que pretenden representar el problema bajo análisis evadiendo los aspectos de incertidumbre. Dentro de estos modelos podríamos considerar la programación lineal, algoritmos de transporte y asignación, algunos modelos de inventarios y la programación dinámica.

En cualquier situación, la toma de decisiones se realiza en un ambiente en el que el decisor no puede estar seguro del comportamiento futuro de ciertos factores que están fuera de su control pero que eventualmente afectarán los resultados de su decisión. Por ejemplo, los resultados de la decisión de explorar y explotar un lote petrolero en la selva peruana se verán afectados por una serie de factores inciertos, como el tamaño de reservas existente, los costos de desarrollo y operación del lote escogido y el precio del petróleo en el mercado internacional. Por ser todos estos factores elementos importantes que afectan la decisión de invertir y el tipo de contrato a negociarse, deberán incluirse explícitamente en el análisis. La teoría de decisiones usa el lenguaje de probabilidades para describir y cuantificar tales incertidumbres y las incorpora en un "árbol de decisiones". Este diagrama permite estructurar el proceso de toma de decisiones bajo incertidumbre. La teoría de decisiones se presenta en el capítulo VIII.

I. Técnicas descriptivas

1. Las variables y sus tipos. 2. Tipos de análisis estadístico. 3. Técnicas de la estadística descriptiva. 4. Técnicas gráficas. 5. Medidas de tendencia central o promedios. 6. Medidas de dispersión o variabilidad. 7. Otras medidas descriptivas: Asimetría y curtosis.

Cualquiera sea el entorno donde se lleve a cabo el proceso de toma de decisiones, es crucial contar con toda la información posible sobre las características relevantes del medio donde se realizan las actividades específicas. La cantidad de información que se genera y recopila ha crecido en forma extraordinaria en los últimos años. La habilidad para absorber, interpretar y utilizar adecuadamente esta información ha adquirido importancia creciente en todas las áreas del quehacer humano.

La evolución de las computadoras ha contribuido sustancialmente a la tendencia de generar gran cantidad de información. El desarrollo alcanzado en computación facilita cada vez más las tareas de almacenar, recuperar, procesar y analizar la información en cantidades insospechables hasta hace unas décadas.

Estamos forzados, cada vez más, a pensar en los aspectos numéricos de cualquier problema de nuestro interés. Existe la necesidad de resumir y presentar la información numérica en una forma tal que permita obtener una apreciación inteligente de lo que significan estos datos. Esta tarea no es de ninguna

manera fácil. Por ejemplo, si tuviéramos acceso a los datos sobre los niveles de ingreso de los dos millones de familias que residen en Lima Metropolitana, ¿qué podríamos hacer con esta información? Concentrarnos en analizar cada uno de los ingresos familiares nos produciría sólo un dolor de cabeza. La única manera de aprovechar esta vasta información es extrayendo la esencia de los dos millones de datos en la forma más directa y simple posible. Uno de los objetivos del análisis estadístico es resumir las características importantes de los datos de tal manera que se obtenga una imagen clara y precisa, sin sacrificar u ocultar características importantes de la información.

El objetivo de este capítulo es revisar algunos de los métodos empleados en el resumen y presentación de información numérica. Esta presentación de la información se logra a través de las tablas estadísticas y gráficas, como veremos en las secciones 3 y 4. Podemos describir los datos en una forma aun más concisa usando técnicas numéricas que nos permiten calcular promedios, medidas de dispersión, simetría y curtosis. Estas técnicas se presentan en las tres últimas secciones de este capítulo.

1. LAS VARIABLES Y SUS TIPOS

El análisis estadístico tiene como fin observar y estudiar las *características* de los datos numéricos. Las observaciones pueden referirse a personas, organizaciones, naciones, familias, objetos, etcétera.

Si una *característica* de nuestro interés puede tomar *distintos valores* o tiene *diferentes resultados*, se llama *variable*. Es decir, una variable es una característica de las observaciones que puede ser clasificada por lo menos en dos categorías. Y el *valor* de una variable es una de las categorías en las cuales la variable puede ser clasificada. A continuación se presentan algunos ejemplos de variables y posible conjunto de valores que estas pueden tomar.

Nombre de la variable	Valores
Sexo (S)	Masculino, femenino
Profesión (P)	Ingeniero, médico, otra
Estatura(E)	Pequeño, mediano, alto
Estatura en centímetros (E')	0,1,2,3,4,5...
Edad (X)	$0 < X < \infty$
Tamaño familiar (Y)	$Y = 0,1,2,3,4,5...$

Es posible definir una variable en diferentes maneras de acuerdo con las necesidades del análisis. Por ejemplo, la variable profesión (P) puede clasificarse en muchas más categorías desagregando el valor "otra" e incluyendo contador, abogado, etcétera. Las variables E y E' ilustran dos maneras en que se puede definir la estatura de una persona. La notación $Y = 0,1,2,3,4,5...$ indica que la variable tamaño familiar puede tomar un valor entero positivo. Los puntos suspensivos indican "y así sucesivamente". La variable X es la edad de una persona medida por el tiempo real que ha transcurrido desde su nacimiento (en años, meses, días, horas, etcétera). Así, el valor X de una persona aumenta continuamente a medida que pasa el tiempo. La notación $0 < X < \infty$ indica que X puede ser cualquier número no negativo, aunque sabemos que es muy difícil que una persona viva más de 100 años.

El procedimiento a través del cual se determina el valor que se le asigna a una variable se denomina *medición*. Podemos medir la estatura de una persona preguntándole si es pequeña, mediana o alta, o a través de un procedimiento convencional para medir su estatura en centímetros. El valor registrado de una medida simple se llama un *valor observado*. El conjunto de valores observados se denota como *datos* observados.

Las variables pueden clasificarse en dos grandes tipos: variables cualitativas y variables cuantitativas.

Las *variables cualitativas* son las que no pueden expresarse numéricamente; tienen la naturaleza de categoría o clase. Por ejemplo el sexo, la profesión, el lugar de nacimiento.

Las *variables cuantitativas* son aquellas que sí pueden expresarse numéricamente. Si una variable cuantitativa sólo puede tomar un número finito de valores posibles en un cierto rango de valores, se dice que es una *variable discreta*. Si, en cambio, una variable puede tomar un número infinito de valores posibles en un rango dado, se dice que es una *variable continua*. En otras palabras, podemos decir que existen espacios vacíos entre los posibles valores de una variable discreta, pero que los valores de una variable continua entre dos puntos dados pueden representarse a través de una línea continua entre esos dos puntos. Ejemplos de variables discretas son: tamaño familiar, ventas de televisores en una semana, número de accidentes. Ejemplos de variables continuas son estatura, edad, temperatura.

Algunas variables continuas pueden ser definidas como tales o como variables discretas dependiendo de cómo se describen y se utilizan. Por ejemplo, la variable temperatura puede ser considerada como variable discreta si la definimos como la temperatura al grado más próximo.

2. TIPOS DE ANÁLISIS ESTADÍSTICO

Como se mencionó anteriormente, mediante el análisis estadístico se observan y estudian las características de los datos para ubicar y resolver problemas, identificar oportunidades, hacer seguimiento de desempeño y, en general, para facilitar la toma de decisiones.

Existen dos tipos de análisis estadístico: la estadística descriptiva y la estadística inferencial. La *estadística descriptiva* consiste en describir las características de un grupo particular de personas, cosas o fenómenos sin especular acerca de sus características en el futuro bajo diferentes condiciones. En otras palabras, a través de la estadística descriptiva pretendemos conocer el todo, universo o población del cual provienen nuestras observaciones. El objetivo de este análisis es extraer de los datos algunas propiedades que *describan* adecuadamente la estructura del pro-

ceso y el comportamiento del sistema bajo estudio. Por ejemplo, si deseamos conocer el ingreso familiar de los estudiantes de la Escuela de Postgrado de la Universidad del Pacífico para establecer la política de pensiones, nuestro interés no estará en el listado de los 271 niveles de ingreso de cada uno de los estudiantes. Sería de mayor utilidad resumir dichos números a ciertas medidas de ingreso promedio y de variación de dichos ingresos. En el resto de este capítulo discutiremos los diferentes procedimientos y técnicas de la estadística descriptiva.

La *estadística inferencial* consiste en el proceso de hacer generalizaciones o predicciones con base en información limitada o muestral. Con este tipo de análisis formulamos y probamos supuestos (hipótesis) sobre los que establecemos afirmaciones que muestran cuán seguros estamos de que un evento proyectado ocurra. Por ejemplo, si queremos medir el ingreso promedio de los estudiantes de postgrado, podemos tomar una muestra de ellos, y, sobre la base de dicha información, estimar el ingreso promedio de los 271 estudiantes. De igual manera, si deseamos estimar el consumo de energía en Lima en el futuro, podríamos relacionar el consumo de energía de los años pasados con la producción industrial de esos años. Luego estimamos los niveles de consumo de energía que pueden ser proyectados para cada nivel de producción industrial en el futuro, lo que facilitará la toma de decisiones en el sector. Los métodos para hacer inferencias son detallados en los siguientes capítulos del libro.

3. TÉCNICAS DE LA ESTADÍSTICA DESCRIPTIVA

Para poder utilizar la información estadística es necesario, primero, organizarla y resumirla. En la presentación de los datos es importante ser tan conciso como sea posible, pero sin dejar de considerar información esencial. El usuario de información no desea ver cada una de las observaciones individuales; por el contrario, logrará un mejor entendimiento de la información si es que esta se presenta en una forma sucinta. Una manera simple

y conveniente de resumir los datos es tabulándolos. La *tabulación* consiste en agrupar observaciones similares en clases y en mostrar un resumen de cada grupo. Esto se presenta en *tablas estadísticas,* que incluyen títulos, encabezamientos y notas explicativas para que se pueda entender el significado de la información presentada.

Supongamos que deseamos tener información sobre la edad de los estudiantes de la Escuela de Postgrado de la Universidad del Pacífico para evaluar en cierta medida el grado de experiencia que poseen. De los registros de la Escuela podemos obtener los datos de cada uno de los 271 estudiantes inscritos.

Nombre	Edad
Aguirre García, Juan Manuel	35
Álvarez Torres, Mario Alejandro	41
Antúnez Pérez, Ana Marcela	28
⋮	

La lista obtenida del registro aparece en orden alfabético y es completa, pero está tan desagregada que no es muy útil. Una mejor forma de presentación implicaría la omisión de los nombres, y la tabulación de las edades en una tabla estadística de frecuencias o *distribución de frecuencias,* como se muestra en el cuadro 1. En esta tabla de frecuencias se muestra el número de elementos que hay en cada *clase.* En nuestro ejemplo cada clase la constituye un año de edad. De esta tabla podemos determinar cuántos estudiantes tienen 23, 24, 25, ... años, y con esta información podemos tener una idea de la experiencia profesional que poseen.

Aunque el cuadro 1 presenta la información de las edades de los estudiantes en forma tabulada, es posible agregar aún más datos para facilitar el análisis. Esto se logra haciendo un esquema para agrupar las edades en clases más amplias.

CUADRO 1: TABLA DE FRECUENCIAS, EDAD DE LOS ESTUDIANTES DE POSTGRADO
(Matriculados en el semestre 89-3)

Edad	Nº de estudiantes (Frecuencia)
23	5
24	1
25	14
26	14
27	20
28	19
29	25
30	28
31	18
32	22
33	29
34	16
35	8
36	5
37	14
38	7
39	7
40	4
41	3
42	4
43	1
45	1
46	1
47	2
53	1
54	1
60	1
Total	271

La definición de las clases siempre tiene elementos de arbitrariedad, pero es posible establecer algunas recomendaciones que deben tomarse en cuenta cuando se definen las clases para una distribución de frecuencias.

En primer lugar, el *número de clases* a establecerse no debe ser tan grande como para evitar el objetivo de resumir la información, pero tampoco tan pequeño que dé como resultado una pérdida excesiva de información. Como recomendación general, el número de clases debe estar entre cinco y veinte, dependiendo de la naturaleza de las observaciones.

En segundo lugar, es conveniente que el *tamaño de las clases* sea el mismo para todas y cada una de ellas. Este tamaño se calcula como la diferencia entre los valores máximo y mínimo, dividida entre el número de clases. Sin embargo, hay situaciones en las cuales es impráctico tratar de cumplir estrictamente esta recomendación. Cuando se tenga un rango de valores muy amplio será necesario recurrir a clases abiertas, en uno o ambos extremos de la distribución. Al usar clases abiertas se elimina la presencia de clases con poca o ninguna frecuencia.

En tercer lugar, los *límites de las clases* deben determinarse de tal manera que cada observación sea incluida en una y sólo una clase.

Siguiendo las recomendaciones anteriores, podemos construir una tabla de frecuencias por clases para nuestro ejemplo de las edades de los estudiantes, como la que se presenta en el cuadro 2. Nótese que las últimas tres clases pueden agruparse en una clase abierta cuyo intervalo será "48 ó más", con una frecuencia de 3 estudiantes, pero cuya marca de clase no queda definida.

CUADRO 2: TABLA DE FRECUENCIAS POR CLASES
(Edades de los estudiantes de la Escuela de Postgrado)

Clases (Años)	Marca de clase (Años)	Número de estudiantes (Frecuencia)
23 - 27	25	54
28 - 32	30	112
33 - 37	35	72
38 - 42	40	25
43 - 47	45	5
48 - 52	50	0
53 - 57	55	2
58 - 62	60	1
Total		271

Las tablas estadísticas que agrupan los datos en clases servirán para introducir las diferentes *técnicas descriptivas*. Ellas pueden ser clasificadas en dos grandes categorías:

a. *Técnicas gráficas*. Como su nombre lo indica, destacan la presentación gráfica de los datos.

b. *Técnicas numéricas*, las cuales enfatizan el resumen de la información. Estas técnicas generan cuatro tipos de medidas para describir los datos en unos cuantos números:

(i)Medidas de tendencia central o promedios

(ii)Medidas de dispersión o variabilidad

(iii)Medidas de asimetría

(iv)Medida de agudeza o curtosis

En la próxima sección se presentan los principales tipos de métodos gráficos. En la sección 6 se exponen las medidas de tendencia central, que permiten identificar el centro de la distribución. La sección 7 introduce las medidas de dispersión, que complementan la descripción que ofrecen los promedios al me-

dir cuán típicos son estos. Finalmente, en la sección 8 se presentan las otras dos medidas descriptivas: asimetría y curtosis.

4. TÉCNICAS GRÁFICAS

La descripción de un conjunto de datos puede visualizarse más fácilmente a través de un gráfico. A pesar de que existe una gran variedad de gráficos para presentar una serie de datos, los más útiles son los histogramas y los diagramas circulares.

Los *histogramas* pueden usarse para representar tanto variables cualitativas como cuantitativas. El histograma del gráfico 1 representa el número de estudiantes de postgrado según profesiones. Las profesiones, por ser una variable cualitativa, no tienen un orden establecido, y su posición en el eje de las abscisas o X no tiene ningún significado cuantitativo. La altura de las barras representa el número de estudiantes de cada profesión. El gráfico 2 es la representación gráfica de la distribución de frecuencias de las edades de los estudiantes expuesta en el cuadro 2. Todas las posibles edades se registran en el eje de las abscisas en forma ascendente, y su frecuencia en el eje de las ordenadas.

En los histogramas se pueden graficar las frecuencias absolutas observadas o las frecuencias relativas. Estas últimas se calculan dividiendo las frecuencias absolutas por el número total de observaciones. La escala de estas gráficas debe escogerse cuidadosamente, de modo que permita al lector interpretar fácilmente la altura de las barras.

El gráfico 3 muestra el histograma para la distribución de frecuencias por clases del cuadro 2, con la diferencia que las últimas tres clases han sido agrupadas en una clase abierta. Las marcas de clase se han señalado en el eje de las X. La frecuencia correspondiente a cualquier clase se representa por la altura de la barra cuya base es la clase en cuestión.

Gráfico 1: Número de estudiantes, por profesión
(Matriculados en el semestre 89-3)

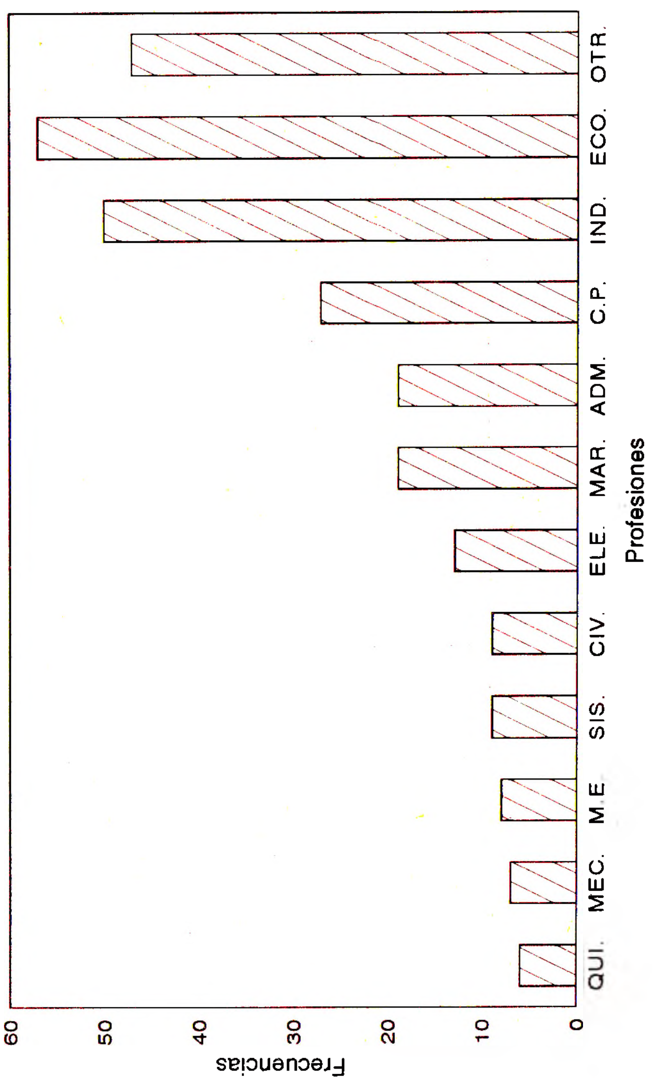


Gráfico 2: Frecuencias de edades de estudiantes de postgrado
(Matriculados en el semestre 89-3)

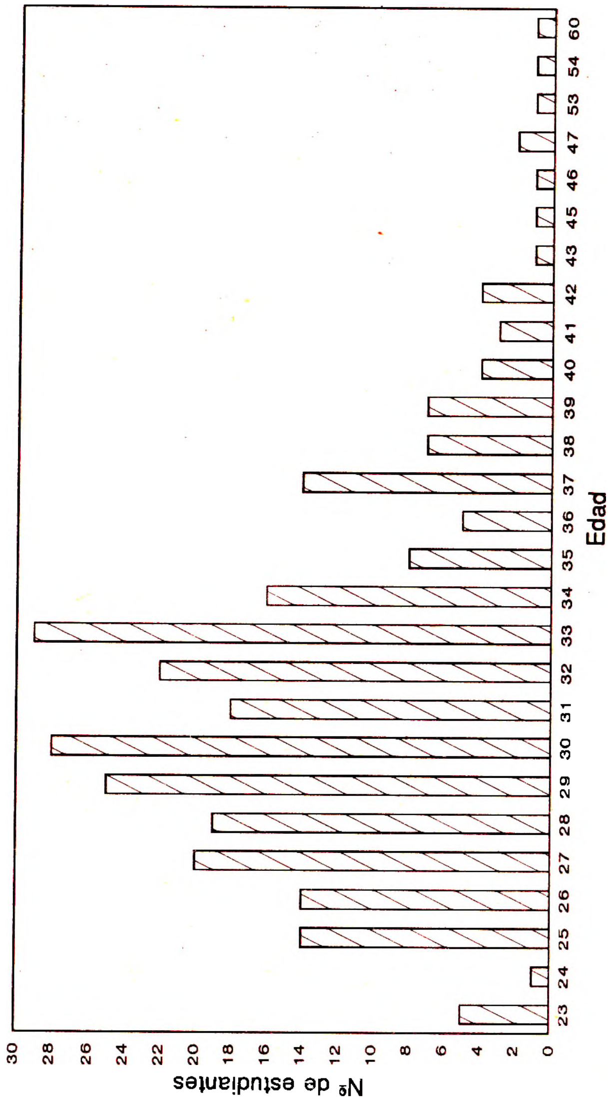


Gráfico 3: Distribución por rango de edades
(Matriculados en el semestre 89-3)



Otro tipo de gráfico útil, sobre todo para representar variables cualitativas, son los *diagramas circulares*. Estos diagramas se usan para representar la división de un todo en sus componentes. Para ilustrar la construcción de este diagrama usaremos el ejemplo de la distribución de los estudiantes según profesiones. El cuadro 3 muestra las frecuencias absolutas, relativas y finalmente las frecuencias relativas transformadas en grados, para poder dividir el diagrama circular convenientemente.

CUADRO 3: TABLA DE FRECUENCIAS RELATIVAS SEGÚN PROFESIONES DE LOS ESTUDIANTES DE POSTGRADO
(Matriculados en el semestre 89-3)

Profesión	Frecuencias absolutas	Frecuencias relativas (%)	Frecuencias relativas (Grados)
Ingeniero químico	6	2.21	7.97
Ingeniero mecánico	7	2.58	9.30
Mecánico electricista	8	2.95	10.63
Ingeniero de sistemas	9	3.32	11.96
Ingeniero civil	9	3.32	11.96
Ingeniero electricista	13	4.80	17.27
Marino	19	7.01	25.24
Administrador	19	7.01	25.24
Contador	27	9.96	35.87
Ingeniero industrial	50	18.45	66.42
Economista	57	21.03	75.72
Otros	47	17.34	62.44
Total	271	100.00	360.00

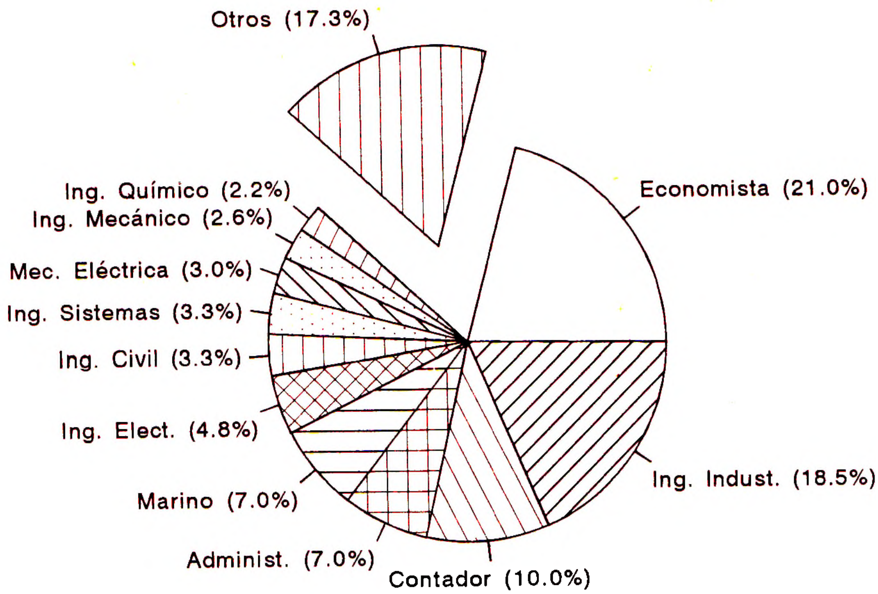
El gráfico 4 muestra la información del cuadro 3 representada en un diagrama circular. El área del círculo representa el número total de alumnos, y los segmentos cortados a partir del centro del círculo expresan la participación que cada profesión tiene del total. El diagrama circular se construye de tal manera que el área de cada segmento es proporcional al número de estudiantes de la profesión correspondiente. Así, por ejemplo, los 27 estudiantes que son contadores representan una proporción de 27/271 = 0.0996 del total de estudiantes. Luego, el área del segmento que corresponde a los estudiantes contadores es proporcional al 9.96 % del área total del círculo. El diagrama circular es fácil de dibujar manualmente. Sabemos que el círculo tiene un total de 360 grados, y la frecuencia total relativa del número de estudiantes es 100%. Por lo tanto, el 1% es equivalente a 3.6 grados. Luego, el segmento que representa al número de estudiantes que son contadores tiene un ángulo que equivale a 9.96 * 3.6 = 35.87 grados.

5. MEDIDAS DE TENDENCIA CENTRAL O PROMEDIOS

Como ya se dijo, para poder analizar adecuadamente la información estadística es necesario que esta se organice y resuma. El agrupamiento de la información en clases tiene la ventaja de presentar la serie de observaciones en una forma más compacta. Sin embargo, para muchas aplicaciones la distribución de frecuencias resulta todavía muy difusa, y nos gustaría contar con un solo valor que represente el orden general de magnitud de los datos observados. A ese número se le conoce como *promedio*, el cual permite condensar la información utilizando un solo valor para describir la distribución.

Un promedio es un valor que debe identificar el centro de una distribución. Por eso, los promedios también se llaman *medidas de tendencia central*. Hay diferentes medidas de tendencia central que pueden usarse para describir los datos, y su selección de-

Gráfico 4: Número de estudiantes según profesión
(Matriculados en el semestre 89-3)



penderá de la naturaleza de los datos y de las necesidades de análisis del usuario. En esta sección se presentan los promedios principales: la media aritmética, la mediana y la moda. Además, definiremos otros promedios de menor uso: la media armónica y la media geométrica.

A. Media aritmética

La media aritmética (o simplemente media) de un conjunto de datos es la suma de todos los valores observados dividida entre el número de observaciones. Si denotamos con la letra *x* a la variable cuyos valores queremos promediar, la media aritmética de una serie de valores de dicha variable (x_1, x_2, x_3,..., x_n) está definida por la siguiente expresión:

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N} = \frac{1}{N}\Sigma x_i \qquad (1)$$

donde μ es la media aritmética
x_i es el valor i-ésimo de la variable x
Σ es la sumatoria de los valores de x_i de i=1 a i=N
N es el número total de observaciones de la variable x

Nótese que el subscrito i es un contador que identifica el primero valor de x cuando i=1, el segundo cuando i=2, y así sucesivamente.

A la media aritmética también se le conoce como *promedio común* o simplemente *media*. A menudo se elimina el subscrito i de la fórmula (1), quedando la media definida de una manera más simple:

$$\mu = \frac{\Sigma x}{N}$$

La media aritmética de las edades de los estudiantes de la Escuela de Postgrado está dada por la suma de todas las 271 edades dividida entre 271.

$$\mu = \frac{23 + 23 + 23 + ... + 53 + 54 + 60}{271} = 32.44 \text{ años}$$

a. Cálculo de la media de una distribución de frecuencias

La media aritmética de una serie de valores puede calcularse a partir de su tabla de frecuencia, mediante la siguiente fórmula:

$$\mu_f = \frac{\Sigma f_i m_i}{\Sigma f_i} \qquad (2)$$

donde μ_f es la media de una distribución de frecuencias
f_i es la frecuencia de la i-ésima clase
m_i es el punto medio (o marca) de la i-ésima clase
k es el número de clases

Debemos notar que $\Sigma f_i = N$.

Es decir, que para calcular la media de una distribución de frecuencias se multiplican las frecuencias de cada clase (f_i) por el punto medio de cada clase (m_i); después se suman estos productos, y el resultado se divide entre la suma de todas las frecuencias.

Dado que la agrupación de los datos en clases implica la pérdida de información precisa de las observaciones individuales, el cálculo de la media sobre la base de la tabla de frecuencias sólo será una aproximación. En el cálculo de esa media se supone que el valor promedio para todas las observaciones de una clase coincide con el punto medio de la clase.

La media calculada con base en la distribución de frecuencias resulta una mejor aproximación del valor verdadero de la media cuando los intervalos de clase son de una amplitud relativamente pequeña.

En el caso de las edades de los estudiantes, si usamos la distribución de frecuencias del cuadro 1, donde el intervalo de clase es un año, la media calculada usando la fórmula (2) coincide con aquella calculada con base en las 271 observaciones individuales (32.44 años). En cambio, si utilizamos la distribución de frecuencias del cuadro 1, donde la amplitud de las clases es de cinco años, obtendremos $\mu_f = 31.83$ años.

b. Cálculo de la media aritmética ponderada

El concepto de media ponderada es una extensión del concepto de media aritmética. En el cálculo de esta última, cada observación se incluye una sola vez en la suma total. En cambio, una *media ponderada* es aquella en la que cada dato u observación es afectada por un factor de ponderación.

$$\mu_p = \frac{\sum_{i=1}^{k} p_i x_i}{\sum_{i=1}^{k} p_i}$$

donde	μ_p	es la media ponderada
	p_i	son los pesos o factores de ponderación que se asigna a cada valor de x_i
	x_i	son los valores observados
	k	es el número de valores observados

La media no ponderada puede definirse como un promedio ponderado donde a cada observación se le da la misma ponderación ($p_i = 1$ para todas las i); entonces, la suma de las ponde-

raciones será igual al total de observaciones ($\Sigma p_i=N$), y la ecuación (3) se convierte en la ecuación (1).

Consideremos el siguiente ejemplo: En el curso de Métodos Cuantitativos I, el sistema de calificación asigna una ponderación de 4 al examen final, 3 al examen parcial, 2 a los controles de lectura y 1 a las tareas. Estas diferentes ponderaciones han sido determinadas considerando que el examen final es más difícil que el parcial, y que el examen parcial es más importante que los controles de lectura y las tareas. Supongamos que un estudiante obtiene una nota de 14 en el examen final, 16 en el examen parcial, 13 en los controles de lectura y 18 en las tareas. Entonces, el promedio ponderado de las calificaciones obtenidas –y, por lo tanto, la nota final– estará dado por:

$$\mu = \frac{4*14 + 3*16 + 2*13 + 1*18}{4 + 3 + 2 + 1} = 14.8$$

B. Mediana (Md)

La mediana es el valor que está en el centro de las observaciones ordenadas tomando como base sus valores, de mayor a menor o viceversa. Si el número de observaciones es par, entonces no existe un valor en el medio, y la mediana se calcula como la media aritmética de los dos valores centrales. En otras palabras, la mediana es aquel punto o posición en la escala de la variable que divide al ordenamiento en dos partes iguales.

En el ejemplo de las edades de los estudiantes de la Escuela de Postgrado, para calcular la mediana primero debemos ordenar los datos en forma ascendente. Así, la mediana se define por el valor que está en la posición 137, esto es, 31 años. Luego, usando la mediana como valor central o típico, la edad promedio será menor que si se usa la media aritmética (μ =32.44). Esto se debe a la influencia que los valores extremos superiores de la distribución tienen sobre la media aritmética (ver cuadro 1).

C. Moda (Mo)

La moda es aquel valor de la distribución que se repite con mayor frecuencia. Es decir, la moda es el valor más común o predominante (el valor que está de moda), dado que representa a más observaciones que cualquier otro valor. Sin embargo, existen distribuciones en las cuales este promedio no es único, pues la mayor frecuencia es compartida por dos o más valores.

Si los datos están agrupados en clases, la moda puede definirse como el punto medio de la clase con la mayor frecuencia. Del cuadro 1 vemos que la distribución de las edades de los estudiantes de postgrado es bimodal: 30 y 33 años. Pero si consideramos la distribución de frecuencias del cuadro 2, el punto medio de la clase 28-32, es decir 30 años, puede usarse como una aproximación de la moda.

D. Media geométrica

Otra medida de tendencia central que posee propiedades convenientes para cierta clase de problemas es la media geométrica. Esta se define como la raíz n-ésima del producto de los n valores de la variable

$$G = (x_1 * x_2 * x_3 * \dots * x_N)^{1/N} \qquad (4)$$

La media geométrica es particularmente apropiada cuando la variable X representa cambios porcentuales. Debemos notar que la media geométrica ha de ser igual a cero si alguno de los valores de x es cero, y que será un valor imaginario si existe un número impar de valores negativos entre los datos.

E. Media armónica

La media armónica se calcula promediando los recíprocos de los valores de la variable y determinando el recíproco de este promedio. Así, primero calculamos el promedio de los recíprocos:

$$\frac{1}{N}\sum\frac{1}{x_i}=\frac{1}{H} \qquad (5)$$

Siendo 1/H el recíproco de la media armónica, esta estará definida por:

$$H=\frac{N}{\sum 1/x_i} \qquad (6)$$

A pesar de que este promedio es poco usado, resulta ser una medida de tendencia central apropiada cuando la variable x ya representa promedios por sí misma. Por ejemplo, al calcular la velocidad promedio de un recorrido total cuando se conocen las velocidades medias para los recorridos parciales (x_i), la media armónica de estos valores será el promedio adecuado (ver el problema 7 al final de este capítulo).

Debemos señalar que la media armónica de una serie de valores positivos siempre será menor que la media geométrica, y esta última menor que la media aritmética, a menos que todos los valores sean iguales, en cuyo caso las tres medias coinciden.

En esta sección se han presentado las diferentes medidas de tendencia central. En ocasiones no es fácil determinar cuál de estas medidas usar en diferentes problemas. Algunos promedios son mejores para ciertos propósitos, mientras que otros lo son para otros propósitos.

La media aritmética es el promedio más conocido por la mayoría de las personas. Además, es el más usado debido a la gran variedad de procedimientos estadísticos que han sido desarrollados haciendo uso de ella. La media aritmética se presta a manipulaciones algebraicas posteriores.

Por otro lado, dado que la media aritmética se calcula sobre la base de todas las observaciones, es afectada por todos los valores de la variable. En ciertas ocasiones esto puede resultar en que algunos valores extremos ejerzan demasiada influencia

en el valor de este promedio. Cuando hay valores extremos que "distorsionen" la media es mejor usar la mediana para representar el promedio de los datos. Otro caso en el que se debe usar la mediana es cuando se utiliza la distribución de frecuencias con una clase abierta (por ejemplo, "48 años o más"), dado que el punto medio de tal clase no puede ser definido, lo que imposibilita calcular la media.

La moda es un promedio que se usa poco, dado que no siempre es una medida única; además, al igual que la mediana, no se presta a manipulaciones algebraicas posteriores. Tanto la media armónica como la geométrica se usan muy poco.

El *mejor promedio* para todos los propósitos no existe. La elección depende del objetivo de la persona que está usando el promedio y de la clase de análisis que se está haciendo. En este libro el promedio más usado es la media aritmética.

6. MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Las medidas de dispersión tratan de describir cuán agrupados o alejados están los datos observados de su promedio. Esto quiere decir que la medida de dispersión provee información acerca de cuán típico es el promedio. Entre más dispersas estén las observaciones individuales, mayor es la medida de dispersión, y menos adecuado será cualquier promedio como medida descriptiva del valor típico. Por lo tanto, es importante que después de elegir y calcular un promedio se determine el grado de variación de las observaciones individuales alrededor de dicho promedio. Las medidas de dispersión que se discuten en esta sección son el recorrido o rango, la desviación intercuartiles, la desviación promedio, la varianza y la desviación estándar. También se introduce el concepto de dispersión relativa, el cual es de utilidad para comparar la dispersión entre dos o más distribuciones de variables con diferentes unidades de medida.

A. Recorrido o amplitud del campo de variación (R)

El recorrido es la medida de dispersión más simple, y se define como la diferencia entre los valores máximo y mínimo de los datos:

$$R = x_M - x_m \tag{7}$$

donde x_M es el valor máximo de la variable x
x_m es el valor mínimo de la variable x

La ventaja de esta medida de dispersión es que resulta sencilla de calcular y fácil de entender. Su desventaja principal es que ofrece poca información, pues se basa solamente en los dos valores extremos y deja de lado la dispersión que existe entre las observaciones restantes. Por ejemplo, el recorrido de las edades de los estudiantes de postgrado (ver cuadro 1) es:

$$R = 60 - 23 = 37 \text{ años}$$

Esta medida de dispersión resulta exageradamente influenciada por las edades correspondientes a los tres estudiantes mayores. De omitir estas últimas observaciones, el recorrido se reduciría significativamente (a 24 años).

B. Desviación intercuartiles

Esta medida de dispersión se construye empleando un método similar al que se usó para definir la mediana. La mediana define un valor de x que divide una distribución en dos partes iguales; es decir, la mitad de los datos tiene valores inferiores a la mediana, y la otra mitad tiene valores superiores. Utilizando el mismo concepto, podemos encontrar también otros valores de x: uno que divida la distribución en un punto tal que la cuarta

parte de los datos tengan valores menores que dicho punto, y el otro de manera que una cuarta parte de los datos tengan valores superiores a él. Estas dos medidas, junto con la mediana, constituyen los tres cuartiles de la distribución. Al cuartil más pequeño se le llama primer cuartil (Q_1); al medio, que es la mediana, se le llama segundo cuartil (Q_2 = Md), y al mayor se le denomina tercer cuartil (Q_3).

Puede construirse una medida de dispersión a partir de los cuartiles, tomando la diferencia entre el tercero y el primero y dividiéndola entre 2. Esta medida se llama desviación intercuartiles o recorrido semiintercuartiles, y se calcula con la fórmula siguiente:

$$Q = \frac{Q_3 - Q_1}{2}$$

En el caso de las edades de los estudiantes de postgrado el rango intercuartil (Q_3 - Q_1) es 6 años, y por lo tanto la desviación intercuartil será 3 años.

C. Desviación promedio

Las dos medidas de dispersión presentadas anteriormente son fáciles de calcular y entender, pero no abarcan toda la información disponible de la distribución de las observaciones individuales. Una mejor medida de dispersión es aquella que se basa en el promedio de las desviaciones de cada observación con respecto a la media de la distribución. Sin embargo, este promedio de las desviaciones sería igual a cero, puesto que, por definición de la media, las desviaciones positivas y negativas se cancelan entre sí:

$$\frac{1}{N}\Sigma(x_i - \mu) = \frac{1}{N}\Sigma x_i - \frac{1}{N}\Sigma\,\mu = \mu - \frac{1}{N}N\mu = 0$$

Para salvar esta dificultad se promedian los valores absolutos de las desviaciones de los datos individuales con respecto a la media. Así, la *desviación promedio* (D.P.) se define con la siguiente fórmula:

$$D.P. = \frac{1}{N}\Sigma\ |x_i - \mu| \qquad (8)$$

donde $|x_i - \mu|$ es el valor absoluto de las desviaciones de los datos individuales con respecto a la media; es decir, se calculan las diferencias $(x_i - \mu)$ y se ignora el signo menos en caso que la diferencia resultante sea un número negativo.

Esta medida de dispersión puede calcularse también usando una distribución de frecuencias en la que los datos han sido agrupados en clases. La desviación de todas las observaciones que pertenecen a una clase se halla multiplicando la desviación de su punto medio (m_i) con respecto a la media, por la frecuencia de la clase. La suma de tales productos, omitiendo los signos, es la suma de las desviaciones con respecto a la media ($\Sigma\ |m_i - \mu|\ f_i$). Luego, la desviación promedio (D.P.) se obtiene al dividir esta suma entre el número total de observaciones:

$$D.P. = \frac{1}{N}\Sigma\ |m_i - \mu|\ f_i$$

Debe notarse que esta fórmula involucra la suma de valores absolutos que puede ser difícil de manipular en posteriores aplicaciones.

D. Varianza y desviación estándar

Las medidas de dispersión más utilizadas en el análisis estadístico son la *varianza* y la *desviación estándar o típica*. Su cálculo se realiza tomando como base las desviaciones de todos los valores

observados con respecto a su promedio. La diferencia con la D.P. radica en que las desviaciones son elevadas al cuadrado antes de promediarlas, para evitar la compensación de sus signos. La varianza se define por:

$$\sigma^2 = \frac{1}{N} \Sigma (x_i - \mu)^2 \tag{10}$$

La varianza (σ^2) de un conjunto de N datos es, entonces, definida como la suma de los cuadrados de las desviaciones de las observaciones con respecto a la media, dividida entre el número total de observaciones.

Dado que la varianza es una expresión que involucra valores al cuadrado, para reducir esta medida a las unidades originales de la variable se toma la raíz cuadrada y se obtiene la desviación estándar:

$$\sigma = \sqrt{\sigma^2} \tag{11}$$

E. Dispersión relativa

Las medidas de dispersión absoluta cobran más relevancia cuando las usamos para comparar la variabilidad de dos conjuntos de datos de la misma variable. Supongamos que queremos comparar la dispersión de las edades de los estudiantes de postgrado de la Universidad del Pacífico con las edades de los estudiantes de la Escuela Superior de Administración de Negocios para determinar cuál de estas poblaciones es más homogénea con respecto a su probable experiencia. Esta comparación es factible dado que ambas variables están expresadas en la misma unidad de medida (años). Sin embargo, habrá ocasiones en las que sea necesario comparar la dispersión de dos o más variables expresadas en diferentes unidades de medida. En general, el procedimiento más simple para comparar cantidades expresadas en diferentes unidades de medida es reducirlas a una base

porcentual comparable. En el caso de la desviación estándar podemos eliminar la unidad de medida expresándola como un porcentaje de la media de las observaciones. Esta medida de dispersión relativa se denomina *coeficiente de variación* (V), y se calcula usando la fórmula siguiente:

$$V = \frac{\sigma}{\mu} * 100 \tag{12}$$

7. OTRAS MEDIDAS DESCRIPTIVAS

Para describir los datos hemos usado medidas de tendencia central y medidas de dispersión. Además, existen otras medidas descriptivas menos importantes que sirven para indicar la dirección de la dispersión con respecto al centro de la distribución, y para determinar el grado de "achatamiento" que exhibe la distribución de los datos.

A. Asimetría

Las medidas de dispersión solamente indican la magnitud de las variaciones, pero no proveen información acerca de la dirección hacia la cual se inclina dicha dispersión. La medida de asimetría mide la falta de simetría con respecto al eje vertical de las frecuencias, y muestra la dirección hacia la cual esta se inclina. Si una distribución tiene una cola derecha larga y una izquierda corta, se dice que es asimétrica positiva; en caso contrario, se dice que es asimétrica negativa.

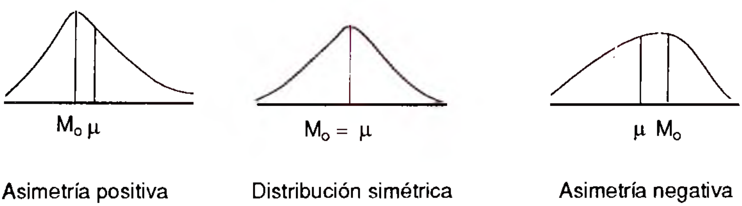
Se puede medir la asimetría de una distribución tomando como base la diferencia entre la media y la moda. Si la distribución es simétrica, la media y la moda son iguales y por tanto la medida de asimetría será cero. Por otro lado, cuanto mayor sea el nivel de asimetría, mayor será la diferencia entre la media y la moda, a causa de los valores extremos. Esta diferencia queda expresada en las unidades de medida de la variable. Para lograr

una medida abstracta, expresamos la diferencia entre la media y la moda en términos de la desviación estándar, definiendo el coeficiente de asimetría:

$$\text{Asimetría} = \frac{\mu - M_0}{\sigma} \tag{13}$$

Usando esta medida de asimetría, vemos que cuando la distribución es asimétrica y se inclina hacia los valores más altos (cola derecha más larga), entonces la media es mayor que la moda y la medida de asimetría será positiva. En cambio, cuando la simetría se inclina hacia los valores más pequeños (cola izquierda más larga), entonces la media será menor que la moda y la medida de asimetría será negativa (ver gráfico 5).

Gráfico 5: Distribuciones con diferentes medidas de asimetría



B. Curtosis

Otra medida descriptiva de una distribución es el grado de "achatamiento" o curtosis que exhibe dicha distribución. Si observamos las curvas del gráfico 5, vemos que todas son simétricas pero tienen diferentes formas. La medida de la magnitud de la curtosis está dada por el "cuarto momento" de la distribución de frecuencias. El cuarto momento es la media aritmética de las desviaciones con respecto a la media elevadas a la cuarta potencia:

$$M_4 = \frac{1}{N} \Sigma (x_i - \mu)^4 \qquad (14)$$

Esta medida es una expresión que involucra valores a la cuarta potencia; para reducir esta medida a un valor abstracto, se la divide entre el cuadrado del segundo momento, que es la varianza al cuadrado.

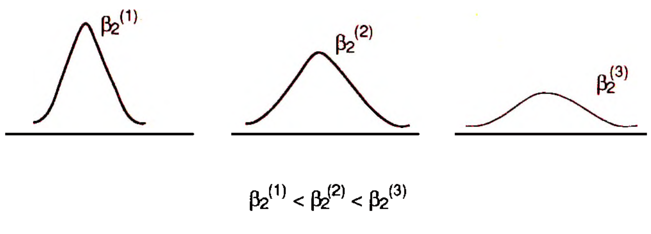
$$M_2 = \frac{1}{N} \Sigma (x_i - \mu)^2 = \sigma^2 \qquad (15)$$

Luego, la medida de curtosis relativa, β_2, se calcula usando la siguiente fórmula:

$$\beta_2 = \frac{M_4}{M_2^2} = \frac{M_4}{(\sigma^2)^2} \qquad (16)$$

Debemos observar que el signo de β_2 es siempre positivo, pues las desviaciones, tanto en el numerador como en el denominador, están elevadas a exponentes pares. El gráfico 6 muestra tres distribuciones simétricas con diferentes grados de "achatamiento".

Gráfico 6: Distribución con diferentes niveles de curtosis



Ejercicios

1. Se ha obtenido datos acerca de 400 "tiempos de lectura" en la biblioteca de la universidad, definidos como el lapso de tiempo que transcurre desde que un estudiante entra en la sala de lectura hasta que sale. Los datos varían de 15 minutos a 3:45 horas. Construya una tabla con diez clases para agrupar estos datos.

2. Los siguientes datos se refieren al tiempo empleado por los trabajadores de una fábrica en movilizarse desde su casa a su centro de trabajo.

Tiempo de transporte	Nº de empleados
Menos de 0.5 hora	60
De 0.5 a 1.0 hora	31
De 1.0 a 1.5 horas	5
De 1.5 a 2.0 horas	3
De 2.0 a 2.5 horas	2
De 2.5 a 3.0 horas	1

a. Trace el histograma de los datos, suponiendo tiempo continuo.
b. Calcule la media aritmética de los tiempos empleados en movilizarse.

3. Los siguientes datos muestran la distribución de frecuencias de los salarios mensuales en dólares de los obreros de la compañía constructora R y G.

Salarios (Dólares)	Nº de empleados
50.00 - 59.99	8
60.00 - 69.99	10
70.00 - 79.99	16
80.00 - 89.99	14
90.00 - 99.99	10
100.00 -109.99	5

a. Construya un histograma para esta variable.
b. Elabore una tabla de frecuencias relativas.
c. Grafique el histograma de frecuencias relativas.
d. Identifique el límite superior de la quinta clase, la marca de clase de la cuarta clase, el tamaño del segundo intervalo de clase, el intervalo de clase que tiene mayor frecuencia, el porcentaje de empleados que tienen un salario menor a $70, el porcentaje de empleados que tienen un salario no mayor a $80.

4. La compañía R y G del problema anterior contrata a cinco nuevos empleados con salarios mensuales de $76.84, $108.76, $122.43, $148.16 y $163.20. Construya:
a. Una distribución de frecuencias para los salarios de los 68 trabajadores.
b. Un histograma para la distribución de frecuencias de a.
c. Un diagrama circular.

5. Las calificaciones de un estudiante del curso de Métodos Cuantitativos son: 15 en el examen parcial, 17 en el examen final, 12 en los controles de lectura y 18 en las tareas. Hallar el promedio de sus calificaciones, si la importancia que se le asigna a estas calificaciones está definida por 30, 35, 20 y 15 respectivamente.

6. Hallar la media aritmética, geométrica y armónica de los siguientes valores: 17, 20, 25, 14, 22, 18 y 25. ¿Qué puede decir de sus valores relativos?

7. Ricardo Carpio viaja de Lima a Arequipa a una velocidad media de 60 kilómetros por hora. El viaje de retorno lo realiza a 80 kilómetros por hora. Hallar la velocidad media para el viaje completo de ida y vuelta. Calcule las medias aritmética y armónica, y comente cuál es la medida correcta de la velocidad media para el viaje completo.

8. Estime el salario promedio para los obreros de la compañía constructora R y G del problema 3.

9. Hallar la media aritmética, el rango, la desviación estándar y el coeficiente de variación para cada uno de los siguientes conjuntos de variables:
a. 8, 9, 8, 9, 8, 3, 9.
b. 5, 18, 10, 15, 3, 7, 6, 12.

Comente acerca de lo adecuado de las diferentes medidas de dispersión en estos casos.

II. Probabilidades

1. Álgebra de eventos. 2. Interpretación de eventos. 3. Árbol de eventos. 4. Fundamentos de probabilidades. 5. Probabilidad condicional e independencia. 6. Expansión en cadena e identidades de expansión. 7. Teorema de Bayes.

El análisis de incertidumbre tiene un papel cada vez más importante en la sociedad moderna. La incertidumbre es una característica fundamental de cualquier entorno en el que se toman las decisiones humanas. En economía, finanzas, administración, ingeniería, medicina y otras disciplinas encontramos problemas que requieren un razonamiento cuantitativo de fenómenos aleatorios. La base de este razonamiento es la teoría de probabilidades. El análisis probabilístico ha evolucionado desde sus orígenes en los juegos de azar del siglo XIV, hasta convertirse en un asunto de interés de todas las personas educadas de hoy.

La teoría de probabilidades no sólo permite analizar incertidumbres, sino también hacer inferencias estadísticas de futuros eventos y acerca de las características de una población basándose en información muestral. La teoría de probabilidades es pues el mecanismo que sirve para lograr uno de los objetivos del análisis estadístico al permitir el uso de información parcial derivada de una muestra para inferir las características de un conjunto mayor de datos que constituyen la población.

Desde varios puntos de vista la teoría de probabilidades es considerada la piedra angular del análisis estadístico. Por lo tanto, el estudiante deberá dedicar especial atención a los conceptos que aquí se exponen. Además, deberá tener cuidado de no cometer el error común de creer que hay algo impreciso o incierto en la solución de un problema aleatorio. Algunos estudiantes piensan que al calcular resultados de un fenómeno aleatorio están utilizando métodos que son de alguna manera imprecisos. Nada más falso. Un problema de probabilidades tiene una respuesta que es tan determinística como la solución de un problema de cálculo. Se obtiene una respuesta por un procedimiento que es tan riguroso y lógico como el de la solución de cualquier problema matemático. El propósito de este capítulo es la revisión de los pasos lógicos entre la formulación del problema y su solución. Para lograrlo se apela a la intuición gráfica del estudiante, eliminando la sensación de misterio que usualmente existe en el estudio del análisis de probabilidades.

1. ÁLGEBRA DE EVENTOS

Cuando se utiliza el lenguaje común para describir acontecimientos o eventos complicados se corre el riesgo de hacerlo con cierta ambigüedad. Para evitar este problema, la teoría de probabilidades recurre a un lenguaje preciso llamado *álgebra de eventos* o *álgebra de conjuntos*.

La *ocurrencia* de un evento se representa con una letra mayúscula *A, B, C*, etcétera. La *no ocurrencia* se representa añadiendo el símbolo (') al evento correspondiente. Así, A' representa la no ocurrencia del evento A; y se le define como el *complemento* del evento A.

La expresión A + B representa la *suma lógica* de los eventos A y B, y se define como la ocurrencia del evento A o del evento B, o de ambos. Se le conoce como la operación de la "*o inclusiva*", y también como la *unión de eventos*, representándola algunas veces con los símbolos $A \cup B$, que se lee "A unión B". Usaremos aquí

el signo más (+) para indicar esta operación. Pero debemos recordar que esta operación tiene propiedades diferentes a la adición aritmética.

La expresión *AB* representa el *producto lógico* de los eventos A y B y se define como la ocurrencia de A y B simultáneamente. Esta expresión también puede representarse como $A \cap B$, "A intersección B". El producto AB es a menudo conocido como la operación "y".

El *evento universal* es el conjunto de todos los eventos que pueden ocurrir, es decir es el evento cierto, denotado por U. Lo opuesto es el evento nulo o *vacío*, es decir, el evento que no puede ocurrir, denotado por Ø.

A continuación se presenta un sumario del álgebra de eventos, basándose en ocho axiomas o relaciones básicas sobre las cuales se pueden construir relaciones más complejas.

Axioma 1	$A + B = B + A$
Axioma 2	$(A')' = A$
Axioma 3	$(A + B) + C = A + (B + C)$
Axioma 4	$(AB)' = A' + B'$; o $AB = (A' + B')'$
Axioma 5	$(A + B)(A + C) = A + (BC)$
Axioma 6	$AA' = \varnothing$
Axioma 7	$A \cup A = A$
Axioma 8	$A + \varnothing = A$

El axioma 1 define la propiedad conmutativa; el axioma 3, la propiedad asociativa, y el axioma 5 la propiedad distributiva.

Estos ocho axiomas son consistentes con la interpretación en términos de eventos. Por ejemplo, el axioma 4, $(AB)' = A' + B'$, expresa que la no ocurrencia del evento AB es equivalente a la no ocurrencia de A o la no ocurrencia de B, o ambas no ocurrencias. El axioma 6, $AA' = \varnothing$, expresa que la ocurrencia de A y de A' es igual al evento nulo, es decir, que A y no A no pueden ocurrir simultáneamente.

A. Diagramas de Venn

Las relaciones entre eventos pueden representarse gráficamente en los llamados diagramas de Venn, como se muestra en el gráfico 7. El evento universal está definido por el área total del rectángulo, mientras que los otros eventos están representados por regiones dentro de dicho rectángulo.

La *suma* o unión de dos eventos es el área contenida en uno de ellos o en ambos; mientras que el *producto* o intersección de dos eventos es el área común a ambos. El complemento del evento A, A′, es el área dentro de U pero fuera de A. El evento nulo, ∅, es el área de dimensión cero.

Los ocho axiomas presentados anteriormente pueden ser graficados en un diagrama de Venn. Por ejemplo, el axioma 2, (A′)′ = A, expresa que el área fuera de A′ pero dentro de U es el área A.

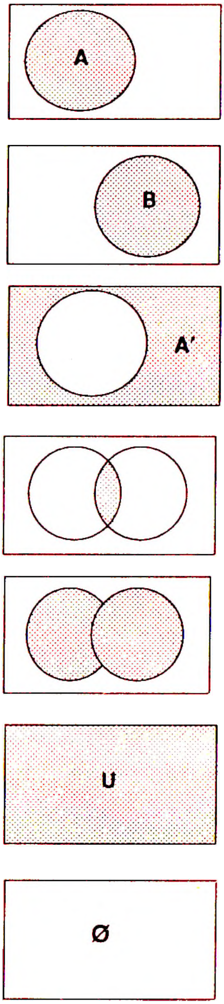
Los axiomas básicos del álgebra de eventos pueden ser extendidos por el proceso usual de derivación. Así por ejemplo, a través de una secuencia de relaciones se prueba que la operación de multiplicación lógica de eventos es conmutativa.

$AB = (A' + B')'$	Usando el axioma 4
$AB = (B' + A')'$	Usando el axioma 1
$AB = BA$	Usando el axioma 4

A través de secuencias similares podemos escribir una lista de relaciones semejantes a los axiomas presentados:

(1′)	AB	=	BA
(2′)	$(AB)C$	=	$A(BC)$
(3′)	$(A + B)'$	=	$A'B'$, o $(A'B')' = A + B$
(4′)	$A(B + C)$	=	$AB + AC$
(5′)	$A + A'$	=	U
(6′)	$A + U$	=	U
(7′)	$A + \varnothing$	=	A
(8′)	$A\varnothing$	=	$\varnothing$

Gráfico 7: Diagramas de Venn



Complemento

INTERSECCIÓN O PRODUCTO

A ∩ B AB

UNIÓN O SUMA

A U B A+B

UNIVERSO U

NULO Ø

Para asegurar un mejor entendimiento del álgebra de eventos, se sugiere que el estudiante intente probar estas últimas relaciones, las cuales pueden ser verificadas inmediatamente utilizando los diagramas de Venn.

Es conveniente derivar algunas otras relaciones para establecer una clara distinción entre el álgebra de eventos y el álgebra usual. Por ejemplo, utilizando el álgebra de eventos:

$$A + A = A$$

Esta relación es fácil de probar usando los axiomas básicos y las relaciones ya derivadas:

$A + AA'$	$=$	$(A + A)(A + A')$	Por el axioma 5
$A + \varnothing$	$=$	$(A + A)\,U$	Por el axioma 6 y (6′)
A	$=$	$A + A$	Por (8′) y el axioma 7

De igual manera, se pueden derivar otras relaciones que enfatizan la distinción entre el álgebra de eventos y el álgebra usual, tales como:

$$A + AB = A$$
$$A + A'B = A + B$$

B. Eventos mutuamente excluyentes

El concepto clave en el álgebra de eventos es el de eventos *mutuamente excluyentes,* que surge cuando la ocurrencia de un evento excluye la ocurrencia de otro u otros. Dos eventos A y B son mutuamente excluyentes si ambos no pueden ocurrir simultáneamente; es decir,

$$AB = \varnothing$$

Si estos eventos son representados en un diagrama de Venn, sus áreas no se sobreponen (ver gráfico 8).

C. Eventos colectivamente exhaustivos

Otro concepto muy útil es el de eventos *colectivamente exhaustivos,* que se define como el conjunto de eventos que poseen la propiedad de que al menos uno debe ocurrir. Así, un conjunto de eventos A_1, A_2, ... A_n son colectivamente exhaustivos si:

$$A_1 + A_2 + ... + A_n = U$$

El gráfico 8 ilustra este concepto.

Finalmente, podemos combinar los conceptos antes mencionados y definir eventos mutuamente excluyentes y colectivamente exhaustivos. Los eventos A_1, $A2$, ... A_n son mutuamente excluyentes y colectivamente exhaustivos si se cumple que:

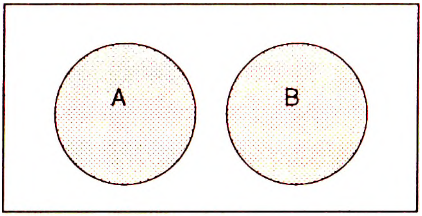
$$A_iA_j = \varnothing \text{ para } i \neq j; \text{ y}$$

$$A_1 + A_2 + ... + A_n = U$$

Por ejemplo, definamos el evento A_i como el resultado de obtener el valor i (i = 1, 2, 3, 4, 5, 6) al lanzar un dado. Los eventos A_1, A_2, A_3, ... , A_6 forman un conjunto de eventos mutuamente excluyentes y colectivamente exhaustivos.

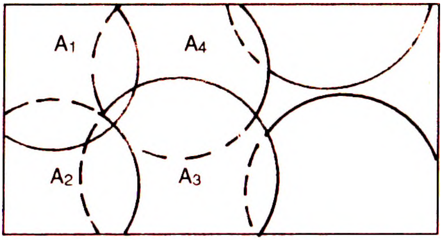
Cualquier expresión de eventos puede re–expresarse en una forma de eventos mutuamente excluyentes, como muestra el gráfico 11 para el evento A + B.

Gráfico 8: Eventos mutuamente excluyentes



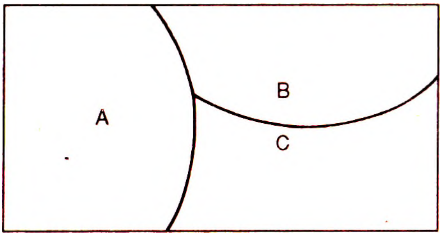
$AB = \emptyset$

Gráfico 9: Eventos colectivamente exhaustivos



$A_1 + A_2 + \ldots + A_n = U$

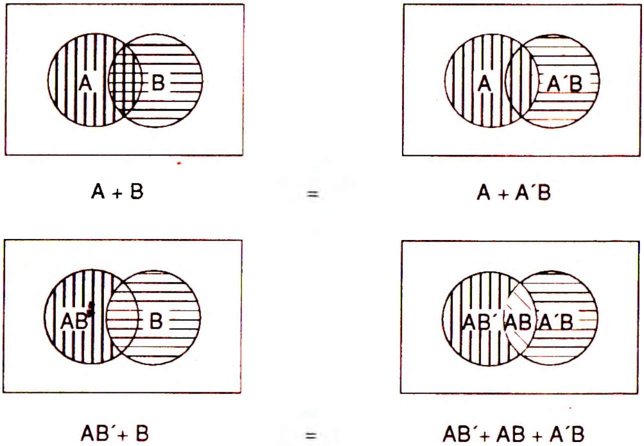
Gráfico 10: Eventos mutuamente excluyentes y colectivamente axhaustivos



$AB = AC = BC = \emptyset$

$A + B + C = U$

Gráfico 11: Formas mutuamente excluyentes de A + B



2. INTERPRETACIÓN DE EVENTOS

El álgebra de eventos es de mucho interés porque nos facilita la solución de problemas que serían difíciles de tratar de otra manera, permitiéndonos evitar ambigüedades inherentes en la comunicación de acontecimientos complicados. Un ejemplo muy simple puede ilustrar la utilidad del álgebra de eventos. Interpretemos el siguiente enunciado:

"Iré a la playa y comeré ceviche o pescado frito."

Esta afirmación podría interpretarse de diferentes maneras. Así, puede significar que quizá coma ambos platos, lo que llamaremos evento E1; o que no coma ambos, evento E2. Con la ayuda del álgebra de eventos podemos precisar el significado de nuestra afirmación. Definamos los siguientes eventos:

P	:	Ir a la playa
C	:	Comer ceviche
F	:	Comer pescado frito

Con estas definiciones se puede precisar el significado de la oración; se comerá "quizá ambos" (E1), o "no ambos" (E2):

$$E1 = P(C + F)$$
$$E2 = P(CF' + C'F)$$

donde F′ es el evento de no comer pescado frito y C′ es el evento de no comer ceviche. Usando el álgebra de eventos podemos expresar los eventos E1 y E2 en una forma de eventos mutuamente excluyentes para establecer la diferencia entre ellos:

E1	=	$P(C + F) = P(CU + FU)$	Por el axioma 7
	=	$P[C(F + F') + F(C + C')]$	Por la relación 6′
	=	$P[CF + CF' + FC + FC']$	Por la relación 5′
	=	$P[CF + CF' + FC']$	Por la relación 1
	=	$P\,CF + P\,CF' + PC'F$	

Mientras que:

E2	=	$P(CF' + C'F)$	
	=	$P\,CF' + PC'F$	Por la relación 5′

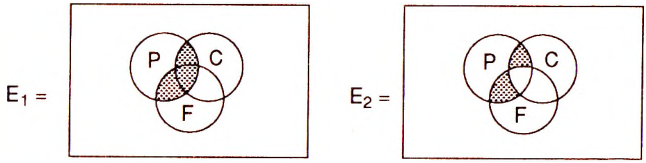
Estos dos eventos se muestran en el gráfico 12, que ilustra la diferencia entre ellos.

3. ÁRBOL DE EVENTOS

Muchas veces resulta conveniente graficar la relación entre eventos en la forma de un *árbol de eventos*. Este árbol es más útil

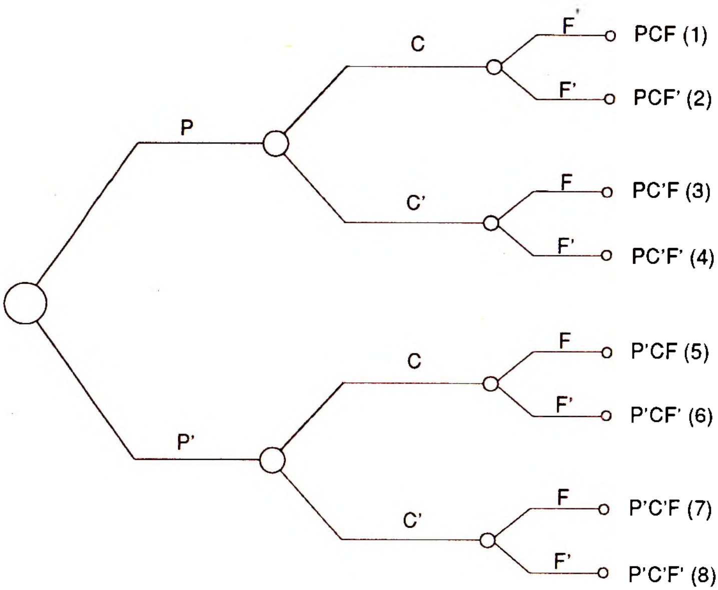
cuando los eventos suceden en forma secuencial y no simultáneamente, pero puede ser usado para cualquier tipo de evento. Además, esta herramienta es apropiada cuando el número de eventos es mayor que tres, y ya no se puede usar el diagrama de Venn para representar la relación entre ellos.

Gráfico 12: Interpretación de eventos



Para construir un árbol de eventos se parte de un punto llamado *nodo* o *nudo*, desde donde se dibuja una línea recta o rama por cada posible resultado de la primera parte del experimento. Se define un experimento usando la terminología de probabilidad, como un proceso que genera resultados que pueden ser definidos de antemano. Cada una de las ramas se denota con el resultado respectivo. Luego usamos los terminales de estas primeras ramas como nodos de comienzo de otras ramas correspondientes a cada posible resultado de la segunda parte del experimento. El proceso continúa hasta que todas las posibles secuencias de resultados hayan sido representadas. Los nodos finales del árbol representan el conjunto de eventos mutuamente excluyentes y colectivamente exhaustivos del experimento. Así, si consideramos el ejemplo anterior, "Iré a la playa y comeré ceviche o pescado frito", tendremos el árbol de eventos del gráfico 13.

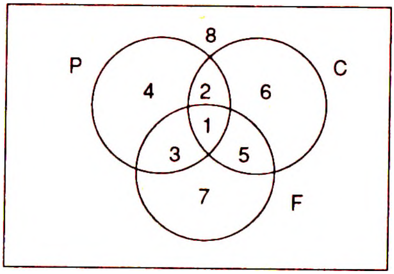
Gráfico 13: Árbol de eventos



Los ocho nodos finales (PCF, PCF', ..., P'C'F') representan todos los resultados posibles de nuestro experimento, que son eventos mutuamente excluyentes y colectivamente exhaustivos, como se aprecia en el diagrama de Venn del gráfico 14.

El árbol de eventos es una herramienta muy útil en la solución de problemas tanto teóricos como prácticos. Sin embargo, para nosotros el álgebra de eventos no es el fin en sí mismo, sino un paso hacia el estudio de probabilidades.

Gráfico 14: Regiones mutuamente excluyentes y colectivamente exhaustivas



4. FUNDAMENTOS DE PROBABILIDADES

En el álgebra de eventos se trató de la ocurrencia o no ocurrencia de un evento. Sin embargo, en la vida real la posibilidad de que un evento suceda puede estar entre lo virtualmente cierto y lo virtualmente imposible. La teoría de probabilidades ofrece un marco para asignar números reales a la posibilidad de ocurrencia de diferentes eventos, de tal manera que sus posibilidades puedan ser comparadas y evaluadas.

Conceptualmente, existen tres maneras de determinar la probabilidad de una ocurrencia: objetiva, experimental y subjetiva.

A. Probabilidad objetiva

Depende de las características físicas del objeto de estudio. Esta manera de asignar probabilidad es conveniente para resolver problemas que involucran el uso de objetos físicos. Así, por ejemplo, la posibilidad de obtener un 4 al lanzar un dado no cargado es 1/6. Como se puede deducir, esta metodología no será apropiada para tratar problemas económicos o administrativos.

B. Probabilidad experimental

Se le llama también frecuencia relativa de una ocurrencia. En algunos casos es posible estimar la probabilidad de ocurrencia de un evento observando el número de veces que ocurrió en un período largo. Así, si se observa que en dos de los últimos cincuenta años ha llovido el día 15 de enero en la ciudad de Lima, esta frecuencia relativa, 2/50, podrá ser utilizada para calcular la probabilidad de que llueva el próximo 15 de enero en Lima.

Esta probabilidad experimental supone que una misma situación pueda repetirse varias veces, y, sobre todo, se presume que tal situación no cambiará en el futuro. Circunstancias de este tipo no son muy frecuentes en las áreas de administración y economía, y por tanto no se pueden calcular probabilidades a través de una serie de repeticiones de un experimento.

C. Probabilidad subjetiva

Es la medida asignada a la valoración subjetiva hecha por un individuo de la probable ocurrencia de un evento. Se basa en la información de que dispone esta persona en un momento dado, es decir en su *estado de información*. Así, por ejemplo, si se desea determinar la probabilidad de que el Perú tenga un gobierno democrático en el año 2000, uno no puede calcular la probabilidad objetiva de ello, ni idear un experimento que proporcione una frecuencia relativa. En estos casos la probabilidad subjetiva es la manera relevante –y en muchas situaciones la única– de asignar probabilidades a una ocurrencia.

La idea fundamental en este tipo de asignación es que la probabilidad es un número que usamos para describir nuestra certeza sobre la ocurrencia de un evento. El grado de certidumbre depende de la información de que disponemos con respecto al evento. Al depender esa medida del estado de información, ella puede cambiar con la disponibilidad de nueva información y puede variar entre diferentes individuos.

Para enfatizar el hecho de que la medida de probabilidad varía con el estado de información, se usará la siguiente notación inferencial:

$p(A/e)$ = Probabilidad del evento A, dado el estado de información e.

La notación equivalente es:

$$p(A/e) = p(A)$$

D. Axiomas de la teoría de probabilidades

La teoría de probabilidades se basa en el álgebra de eventos y en tres axiomas adicionales. La medida de probabilidad es definida como una función real para todos los eventos que pueden ser construidos usando el álgebra de eventos. Los tres axiomas que la función de probabilidades debe satisfacer son:

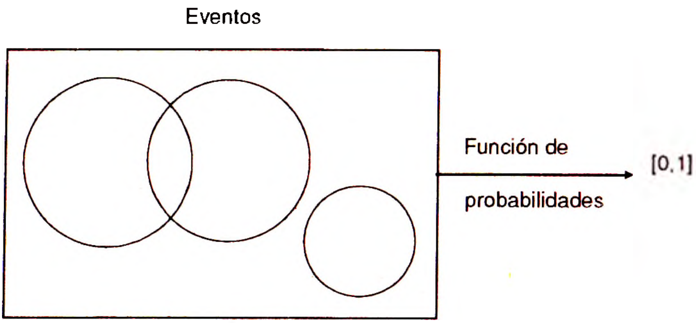
P.1 : $p(A/e) \geq 0$

P.2 : $p(U/e) = 1$

P.3 : Si A y B son mutuamente excluyentes, entonces
$p[(A + B)/e] = p(A/e) + p(B/e)$

El primer axioma (P.1) requiere que la medida de probabilidad de cualquier evento sea un número real no negativo. El segundo normaliza las probabilidades al requerir que la probabilidad del evento cierto sea 1. El tercero requiere que la probabilidad de eventos mutuamente excluyentes sea aditiva. Gráficamente, tenemos:

Gráfico 15: Función de Probabilidades



La función de probabilidades definida por estos tres axiomas es la función necesaria para medir las áreas de cada región en el diagrama de Venn. El axioma (P.1) establece que el área de cualquier región no puede ser negativa; (P.2) expresa que el área total del diagrama debe considerarse igual a 1; (P.3) establece que el área contenida en dos regiones que no se sobreponen debe ser la suma de sus áreas.

Sobre la base de los tres axiomas de probabilidades y los ocho del álgebra de eventos podemos derivar algunas relaciones importantes, como las siguientes:

P.1') $p(A'/e) = 1 - p(A/e)$
P.2') $p(\varnothing/e) = 0$
P.3') $p[(A + B)/e] = p(A/e) + p(B/e) - p(AB/e)$

La relación (P.1') establece que la probabilidad de ocurrencia del complemento de un evento es igual a 1 menos la probabilidad de ese evento. Podemos demostrar esta relación de la siguiente manera:

$$A + A' = U \quad \text{Por la relación (6')}$$
$$AA' = \varnothing \quad \text{Por el axioma A.6}$$

Y usando los axiomas (P.2) y (P.3) podemos escribir:

$$p[(A + A')/e] = p(A/e) + p(A'/e) = 1$$

Y, luego:

$$p(A'/e) = 1 - p(A/e)$$

La probabilidad del evento nulo, la relación (P.2'), se prueba de la siguiente manera. Dado que $U' = \varnothing$, entonces:

$$p(\varnothing/e) = 1 - p(U/e) \qquad \text{De la relación (P.1')}$$

y usando el axioma (P.2),

$$p(\varnothing/e) = O$$

Finalmente, podemos demostrar (P.3'), la probabilidad de la suma de dos eventos que no son mutuamente excluyentes, re–expresando la suma de A y B (ver gráfico 11).

$$A + B = A + A'B \qquad (1)$$

dado que A y (A'B) son mutuamente excluyentes, $A(A'B) = \varnothing$, y, por el axioma (P.3), tenemos:

$$p[(A + B)/e] = p(A/e) + p(A'B/e) \qquad (2)$$

Utilizando las relaciones (A.7), (6') y (5') podemos escribir el evento B en la siguiente forma:

$$B = UB = (A + A')B = AB + A'B \qquad (3)$$

Luego, dado que (AB)(A'B) = Ø, podemos usar otra vez el axioma (P.3) para obtener:

$$\begin{aligned} p(B/e) &= p(AB/e) + p(A'B/e) \quad \text{o} \\ p(A'B/e) &= p(B/e) - p(AB/e) \qquad (4) \end{aligned}$$

Substituyendo (4) en (2) obtenemos:

$$p[(A + B)/e] = p(A/e) + p(B/e) - p(AB/e)$$

Es decir, la probabilidad de la suma de dos eventos es la suma de sus probabilidades menos la probabilidad de su producto. Si los eventos son mutuamente excluyentes, la probabilidad de su producto es cero y la ecuación (P.3') se convierte en el axioma (P.3).

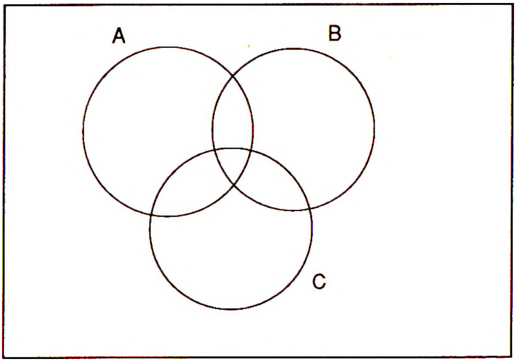
Los diagramas de Venn pueden ser usados para derivar y entender relaciones de este tipo. Así, recordemos que p[(A + B)/e] es el área contenida en la región A o la región B o en la región de A y B. Si calculamos esta área sumando las áreas de las regiones A y B, obtendremos un resultado exagerado porque estaríamos contando el área compartida entre A y B dos veces. Luego, debemos sustraer esta área (AB) una vez para obtener el resultado correcto. Esta operación está establecida en la relación (P.3').

Extensiones de estos argumentos producen otras relaciones importantes. Por ejemplo, suponga que queremos calcular la probabilidad de la suma de tres eventos A, B y C:

$$\begin{aligned} p[(A+B+C)/e] &= p\{[A+(B+C)]/e\} = p(A/e) + \\ &\quad p[(B+C)/e] - p[A(B+C)/e] \\ &= p(A/e) + p(B/e) + p(C/e) - p(BC/e) - \\ &\quad p[(AB+AC)/e] \\ &= p(A/e) + p(B/e) + p(C/e) - p(BC/e) - \\ &\quad p(AB/e) - p(AC/e) + p(ABC/e) \qquad (5) \end{aligned}$$

Esta ecuación tiene una interpretación directa usando el diagrama de Venn del gráfico 16.

Gráfico 16: Diagrama de Venn de tres eventos



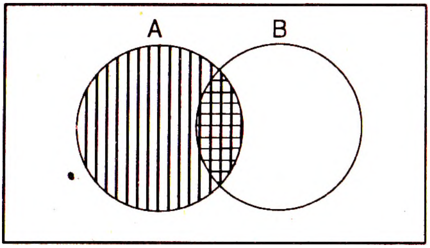
5. PROBABILIDAD CONDICIONAL E INDEPENDENCIA

Una de las definiciones más importantes en la teoría de probabilidades es la de *probabilidad condicional*. Si A y B son dos eventos, y $p(A/e) \neq 0$, entonces "la probabilidad condicional de B dado A", $p(B/A, e)$, se define como el cociente de la probabilidad de que ambos ocurran simultáneamente entre la probabilidad de que el evento condicionador ocurra:

$$p(B/A, e) = \frac{p(AB/e)}{p(A/e)} \qquad (6)$$

Como se aprecia en el diagrama de Venn del gráfico 17, $p(B/A, e)$ es el ratio del área compartida por los dos eventos con respecto al área del evento condicionador.

Gráfico 17: Probabilidad condicional



En cierto sentido, todas las probabilidades son condicionales, porque hemos definido un conjunto de eventos U (conjunto universal) del cual al menos uno ocurre. La probabilidad de cualquiera de estos eventos depende de haber definido el área del conjunto universal 1. Así:

$$p\,(A/U, e) = \frac{p\,(AU/e)}{p\,(U/e)} = \frac{p\,(A/e)}{1} = p\,(A/e) \qquad (7)$$

Si sabemos que el evento A ha ocurrido, entonces A se convierte en el evento cierto, y podemos re–normalizar la probabilidad de cada uno de los otros eventos de tal manera que ellos sumen 1 sobre el área de A. Entonces el conjunto de eventos que se considera que forman el evento universal, U, es arbitrario. La definición de probabilidad condicional nos permite cambiar la base para la asignación de probabilidades en cualquier momento.

Si dos eventos A y B son *independientes*, en el sentido intuitivo, la ocurrencia de A no debería afectar la ocurrencia de B. Es decir:

$$p(B/A, e) = p(B/e), \quad \text{o}$$

$$\frac{p(AB/e)}{p(A/e)} = p(B/e) \qquad (8)$$

Despejando:

$$p(AB/e) = p(A/e)\, p(B/e) \qquad (9)$$

Luego, se define que dos eventos son independientes si la probabilidad de que ambos ocurran simultáneamente es igual al producto de las probabilidades de que ambos ocurran separadamente; o, en otras palabras, dos eventos son independientes si la probabilidad de su producto es igual al producto de sus probabilidades. Entonces usamos la ecuación (9) como la definición formal de independencia. Note que esta definición se basa en la función de probabilidad y no sólo en el álgebra de eventos.

Para enfatizar la definición de independencia, considérese el siguiente experimento que consiste en lanzar una moneda dos veces, y donde la probabilidad de obtener cara es p, siendo $0<p<1$. Definamos los siguientes eventos relacionados al experimento:

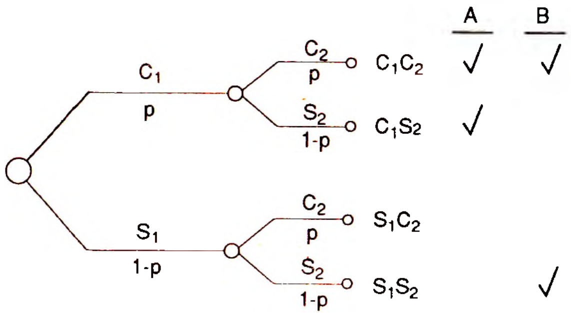
A: "Cara en la primera tirada".
B: "Ambas tiradas dan lo mismo".

Podemos usar el *árbol de probabilidades*, que es un árbol de eventos que incluye información sobre la probabilidad de ocurrencia de cada posible resultado. Es decir, las ramas que denotan los posibles resultados también tienen la información con respecto a la probabilidad de su ocurrencia. Así, el gráfico 18 nos muestra el árbol de probabilidades para este experimento de lanzar la moneda dos veces, y nos indica los eventos A y B.

El evento A es el conjunto de C_1C_2, C_1S_2; el evento B es el conjunto de C_1C_2, S_1S_2; luego, $AB = C_1C_2$. Dado que p es la probabilidad de obtener cara, (1 - p) es la probabilidad de obtener sello. Entonces:

$$p(A/e) = p[(C_1C_2 + C_1S_2)/e] = p(C_1C_2/e) + p(C_1S_2/e)$$
$$= p^2 + p\,(1 - p) = p$$
$$p(B/e) = p[(C_1C_2 + S_1S_2)/e] = p(C_1C_2/e) + p(S_1S_2/e)$$
$$= p^2 + (1 - p)^2$$
$$p(AB/e) = p(C_1C_2/e) = p^2$$

Gráfico 18: Árbol de probabilidades del experimento de lanzar una moneda dos veces



Luego, si p es diferente de 0.5 entonces no se cumple la igualdad $p(AB/e) = p(A/e)p(B/e)$, y los eventos A y B no son independientes. Sin embargo, si $p = 0.5$, entonces $p(AB/S) = p(A/S)\,p(B/S)$ y A y B son independientes. Podemos ver que la cuestión de la independencia puede ser determinada únicamente cuando se conoce la función de probabilidad, y no sólo de la definición de eventos. En cambio, el concepto de eventos mutuamente excluyentes sólo tiene que ver con la definición de los eventos. Así, en el ejemplo vemos que A y B no son mutuamente excluyentes, ya que $AB = C_1C_2 \neq \emptyset$

Los conceptos de eventos mutuamente excluyentes e independientes tienden a confundir a los estudiantes, por lo que se justifica establecer la diferencia entre ellos. En primer lugar,

el concepto de eventos mutuamente excluyentes sólo depende de la definición de eventos ($AB = \emptyset$), mientras que el concepto de independencia depende de la medida de probabilidades [$p(AB/e) = p(A/e)\,p(B/e)$]. En segundo lugar, si tenemos eventos mutuamente excluyentes estos son fuertemente dependientes. Así, si A y B son mutuamente excluyentes y no son ni el evento nulo ni el evento universal ($0 < p(A/e), p(B/e) < 1$), entonces:

$$\begin{aligned} AB &= \emptyset \\ p(AB/e) &= p(\emptyset/e) = 0 \end{aligned} \tag{10}$$

mientras que:

$$p(A/e)p(B/e) \neq 0 \tag{11}$$

Luego, de (10) y (11):

$$p(AB/e) \neq p(A/e)\,p(B/e) \tag{12}$$

Por tanto, los eventos A y B son fuertemente dependientes. Si sucede uno, la probabilidad de que suceda el otro es cero.

Para la manipulación de probabilidades es importante recordar que si se tiene eventos mutuamente excluyentes se pueden sumar sus probabilidades (axioma P.3); mientras que si se tiene eventos independientes se pueden multiplicar sus probabilidades, por la definición de independencia.

6. EXPANSIÓN EN CADENA E IDENTIDADES DE EXPANSIÓN

De la definición de probabilidad condicional, ecuación (6), se deriva el concepto de expansión en cadena expresando el numerador como el producto del denominador por el cociente:

$$p(AB/e) = p(A/e)\, p(B/A, e) \qquad (13)$$

Extensiones de esta relación producen la expansión en cadena para más de dos eventos. Por ejemplo, supongamos que se quiere calcular la probabilidad conjunta de tres eventos ABC:

$$\begin{aligned} p(ABC/e) &= p[A(BC)/e] = p(A/e)\, p[(BC)/A, e] \\ &= p(A/e)\, p(B/A, e)\, p(C/B, A, e) \qquad (14) \end{aligned}$$

Esta ecuación será utilizada más adelante para el cálculo de probabilidades de los resultados de un experimento usando árboles de probabilidades.

Otro concepto importante es el de expansión. Podemos reescribir la probabilidad de ocurrencia de un evento en términos de la ocurrencia de otro evento:

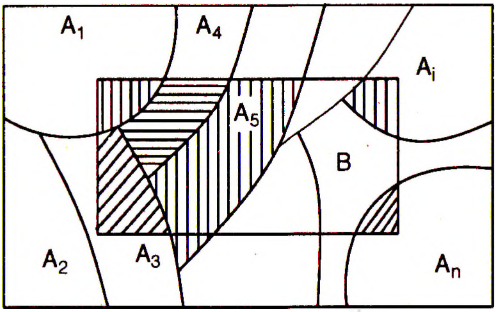
$$\begin{aligned} p(B/e) &= p(B\,U/e) = p[B\,(A + A')/e] = p((BA + BA')/e) \\ &= p(BA/e) + p(BA'/e) \\ &= p(B/A, e)\, p(A/e) + p(B/A', e)\, p(A'/e) \qquad (15) \end{aligned}$$

En general, si consideramos un conjunto de eventos A_1, $A_2,\ldots A_n$ que sean mutuamente excluyentes y colectivamente exhaustivos, entonces:

$$\begin{aligned} p(B/e) &= p(B\,U/e) = p[B(A_1 + A_2 + \ldots + A_n)/e] \\ &= p(BA_1/e) + p(BA_2/e) + \ldots + p(BA_n/e) \\ &= \sum p(B/A_i, e)\, p(A_i/e) \qquad (16) \end{aligned}$$

Esta ecuación tiene una interpretación directa usando el diagrama de Venn del gráfico 19.

Gráfico 19: Expansión del evento B



El área de la región B es igual a la suma de las áreas de las regiones que B comparte con cada una de las regiones A_i, $p(BA_i/e)$, las cuales pueden ser reescritas usando el concepto de expansión en cadena.

7. TEOREMA DE BAYES

El teorema de Bayes establece la relación más importante en la teoría de probabilidades y es la base para la revisión de la asignación de probabilidades a la luz de información adicional. Este teorema se deriva fácilmente de las relaciones básicas expuestas anteriormente.

Consideremos un conjunto de n eventos mutuamente excluyentes y colectivamente exhaustivos $A_1, A_2, \ldots A_n$, y otro evento B, todos con probabilidades de ocurrencia claramente determinadas. Por definición:

$$A_iA_j = \varnothing \qquad 1 \leq i = j \leq n \tag{17}$$

$$A_1 + A_2 + \ldots A_n = U \tag{18}$$

La situación puede ser representada por el diagrama de Venn del gráfico 19. Los eventos A_i dividen el espacio universal, U, en n regiones. El evento B ocupa partes de algunas de esas regiones. Se supone que se conocen las probabilidades de los eventos A_i, y las probabilidades condicionales $p(B/A_i, e)$, para $i = 1, 2, ..., n$.

El problema se plantea con el siguiente interrogante: ¿cuál es la probabilidad de que A_i ocurra si sabemos que B ha ocurrido: $p(A_i/B, e)$? Por definición:

$$p(A_i/B, e) = \frac{p(A_iB/e)}{p(B/e)} \tag{19}$$

Usando el concepto de expansión en cadena escribimos (19) en la forma:

$$p(A_i/B, e) = \frac{p(A_i/e)\, p(B/A_i, e)}{p(B/e)} \tag{20}$$

Esta ecuación expresa la respuesta en términos de las probabilidades conocidas $p(A_i/e)$ y $p(B/A_i, e)$; y el valor de $p(B/e)$ puede ser calculado a partir de ellas usando la ecuación (16).

$$p(B/e) = \sum p(B/A_j, e)\, p(A_j/e)$$

Cuando sustituimos este resultado en (20), obtenemos la forma más común del teorema de Bayes:

$$p(A_i/B, e) = \frac{p(A_i/e)\;\; p(B/A_i, e)}{\sum p(A_j/e)\;\; p(B/A_j, e)} \tag{21}$$

La probabilidad condicional de A_i dado B ha sido expresada en términos de las probabilidades conocidas de los eventos A_i, y de las probabilidades condicionales de B dado A_i. Puesto que $p(A_i/e)\, p(B/A_i, e) = p(A_iB/e)$, podemos escribir la ecuación (21) en una forma más simple:

$$p(A_i/B, e) = \frac{p(A_i/e)}{\sum p(A_j/e)} \quad (22)$$

Esta probabilidad condicional está representada en el diagrama de Venn por el área que A_i comparte con B entre el área total de B, como se aprecia en el gráfico 19.

Hemos dicho anteriormente que todo proceso de toma de decisiones se realiza dentro de los límites que impone el entorno o la naturaleza. El decisor no puede controlar las características del entorno, pero sí está en la capacidad de asignar probabilidades a la ocurrencia de todos los posibles "estados" que estas características pueden tomar. Por ejemplo, si una compañía petrolera decide perforar un pozo, se enfrentará a dos estados de la naturaleza: encontrar petróleo o no encontrar petróleo. Los expertos de la compañía habrán ya calculado las probabilidades de ocurrencia de cada uno de estos "estados de la naturaleza". ¿Qué sucedería si la compañía tiene la oportunidad de realizar un último estudio geológico sobre la presencia de petróleo en el terreno? Los resultados de esta nueva investigación podrían originar la necesidad de revisar las probabilidades de encontrar o no petróleo.

La ecuación (21) es la relación básica para la revisión de probabilidades. Basándonos en la información disponible, podemos interpretar los eventos A_1, A_2, ..., A_n como todos los posibles estados de la naturaleza o del entorno en un contexto específico, y que A_i es el evento en el que la naturaleza está en el estado i, con una probabilidad $p(A_i/e)$. Consideremos que un evento B, estadísticamente relacionado al estado de la naturaleza, ocurre. La pregunta clave es ahora cómo la ocurrencia de B afecta nuestra asignación de probabilidades a los posibles estados de la naturaleza, A_i.

La respuesta está dada por el teorema de Bayes. Llamaremos *probabilidad previa* o *a priori* a la probabilidad original de ocurrencia del evento A_i. Después de observar el evento B, tendremos

una *probabilidad posterior* o *revisada* de ocurrencia de A_i después de ocurrido B, denotada por $p(A_i/B, e)$. Esta probabilidad está definida por el teorema de Bayes y es directamente proporcional a la probabilidad previa multiplicada por lo que se conoce como *función de verosimilitud.* Esta función representa la probabilidad de observar B, si la naturaleza o el entorno estuviera, de hecho, en el estado i. Las probabilidades posteriores son normalizadas con el requerimiento de que ellas sumen 1, como se puede observar en la ecuación (21), en la que el denominador es la suma de todos los posibles valores (i) del numerador.

Ilustraremos la aplicación del teorema de Bayes considerando el caso de la empresa de manufacturas «Excelencia», que compra una pieza o parte básica de dos abastecedores diferentes. Se sabe que el 30% de las piezas compradas por la compañía proviene del abastecedor 1, mientras que el resto son del abastecedor 2. Los eventos A_1 y A_2 definen la procedencia de las piezas del abastecedor 1 y 2 respectivamente. Luego, si seleccionamos aleatoriamente una pieza, tendremos las siguientes probabilidades, *a priori*, de obtener una pieza del abastecedor 1 ó 2:

$$p(A_1/e) = 0.30 \qquad \text{y} \qquad p(A_2/e) = 0.70$$

La calidad de las piezas varía de acuerdo con el abastecedor. Tomando como base los registros históricos, se sabe que la tasa de partes defectuosas del abastecedor 1 es 4%, mientras que la del abastecedor 2 es de 2%. Luego, si D denota el evento de encontrar una pieza defectuosa, entonces tenemos las siguientes probabilidades, que constituyen las funciones de verosimilitud:

$$p(D/A_1, e) = 0.04 \quad \text{y} \quad p(D'/A_1, e) = 0.96$$
$$p(D/A_2, e) = 0.02 \quad \text{y} \quad p(D'/A_2, e) = 0.98$$

Luego, podemos usar el teorema de Bayes para revisar las probabilidades previas a la luz de nueva información. Por ejem-

plo, si sabemos que las piezas adquiridas de ambos abastecedores son usadas indistintamente en el proceso de producción de "Excelencia", y sucede que una máquina se malogra al usar una parte defectuosa, ¿cuál es la probabilidad de que la pieza haya sido adquirida del proveedor 1? Es decir, ¿cuál es la probabilidad posterior $p(A_1/D, e)$? Usando el resultado del teorema de Bayes, ecuación (21), tenemos:

$$p(A_1/D) = \frac{(0.04)\,(0.30)}{(0.04)\,(0.30)\ +\ (0.02)\,(0.70)} = \frac{0.012}{0.026} = 0.46$$

La información adicional de que la parte analizada es defectuosa ha aumentado la probabilidad de que esta pertenezca al abastecedor 1 de 0.30 a 0.46.

Mediante el teorema de Bayes hemos asignado nuevas probabilidades a un conjunto de eventos (A_i) como resultado de observar un evento (D) que está estadísticamente relacionado a dicho conjunto. Para ello necesitamos una distribución de probabilidades *a priori* del conjunto de eventos (A_i) y la función de verosimilitud. Este mecanismo será aplicado en la teoría de decisiones, como veremos en el capítulo VIII.

Ejercicios

1. Sean A = {1,2,3,4,5}, B = {1,3,5,7,9}, donde U = {1,2,3,4,...}. Hallar A + B, AB, A′, B′, A′B′.

2. Definiendo el evento universal U como el conjunto de todos los números enteros positivos, y considerando los siguientes eventos:

$$\begin{aligned} A &= \{X : X \leq 12\} \\ B &= \{X : X < 8\} \\ C &= \{X : X \text{ es par}\} \\ D &= \{X : X \text{ es múltiplo de } 3\} \\ E &= \{X : X \text{ es múltiplo de } 4\} \end{aligned}$$

Exprese, en términos de A, B, C, D y E (y posiblemente sus complementos), los siguientes eventos:
a. {1, 3, 5, 7}
b. {3,6,9}
c. {8, 10}
d. Los números enteros positivos mayores que 12.
e. Los números enteros positivos múltiplos de 6.
f. Los números enteros que son pares y menores o iguales a 6 ó que son impares y mayores que 12.

3. La cadena "Mercados Lima" está considerando incluir tres nuevos productos en su *stock*. Basándose en los resultados de un estudio de mercado que mide la aceptación de estos productos entre los clientes de "Mercados Lima", cada producto será clasificado como "exitoso" o "fracaso".
a. Liste todos los posibles resultados del estudio.
b. Suponiendo que se permite una tercera clasificación para los resultados del estudio para el segundo producto (por ejemplo, "muy exitoso"), dibuje el árbol de eventos para este nuevo experimento. ¿Cuántos nodos finales tendrá el árbol?

4. Cien personas compraron boletos de una lotería. Cuarenta de los cien son mujeres y diez tienen grado de bachiller. Suponiendo que cualquiera de las cien personas tiene la misma probabilidad de ganar la lotería, y que el sexo y la educación universitaria son independientes, encuentre la probabilidad de que el ganador sea:
a. Un hombre.
b. Un hombre sin grado de bachiller.
c. Una mujer con grado de bachiller.

5. El 15% de los sesenta estudiantes graduados en administración en la Escuela de Postgrado de la Universidad del Pacífico en el último año, son mujeres. Una compañía importante solicitó a la Escuela que le recomendara estudiantes graduados para cubrir un puesto vacante. De la lista de todos los graduados se escogió a tres candidatos para una entrevista final. Suponga que estos tres individuos pueden considerarse como una muestra aleatoria de todos los estudiantes graduados en administración.
a. ¿Cuál es la probabilidad de que haya una mujer en la lista?
b. ¿Cuál es la probabilidad de que la mayoría en la lista sean mujeres?
c. ¿Cuál es la probabilidad de que el número de mujeres en la lista esté entre 1 y 2 inclusive?

6. En una entrevista con un ama de casa, se le solicita su opinión sobre cuatro marcas de detergente (A, B, C y D), pidiéndole que indique el orden de su preferencia, marcando con 1 el que prefiere, con 2 el que

le sigue, etcétera. Como la señora en realidad no tiene ninguna preferencia, le asigna a cada marca un número del 1 al 4 completamente al azar.

a. ¿Cuál es la probabilidad de que la marca "A" quede en primer lugar?

b. ¿Cuál es la probabilidad de que "C" quede en primer lugar y "D" en el segundo?

c. ¿Cuál es la probabilidad de que "A" quede en alguno de los dos primeros lugares?

7. En un estudio realizado para investigar la relación entre el hábito de fumar y los ataques cardíacos, se tomó una muestra de 1,000 hombres de más de 50 años de edad, y se obtuvieron los siguientes datos:

180 habían sufrido un ataque cardíaco
300 fueron clasificados como fumadores
100 habían sufrido un ataque cardíaco y fueron clasificados como fumadores
80 habían sufrido un ataque cardíaco y fueron clasificados como no fumadores

a. Dado que un hombre tiene más de 50 años y es un fumador, estime la probabilidad de que sufra un ataque cardíaco, basándose en los resultados del estudio.

b. ¿Cuál es la probabilidad de que un hombre mayor de 50 años que no es fumador, sufra un ataque cardíaco?

c. ¿Demuestra el estudio que los ataques cardíacos y el hábito de fumar son eventos independientes?

8. Los directores de la compañía "Los Gatos Listos" han calculado los siguientes estimados de las probabilidades con respecto a las utilidades anuales del próximo año.

p (Utilidades menores que las de este año) = 0.30
p (Utilidades iguales que las de este año) = 0.50
p (Utilidades mayores que las de este año) = 0.20

Después de mantener conversaciones con los dirigentes sindicales, el Gerente de Personal llegó a las siguientes conclusiones:

p (Sindicato pedirá aumento de salarios/Utilidades menores el próximo año) = 0.25
p (Sindicato pedirá aumento de salarios/Utilidades iguales el próximo año) = 0.40
p (Sindicato pedirá aumento de salarios/Utilidades mayores el próximo año) = 0.90

a. Las probabilidades establecidas por los Directores y el Gerente de Personal, ¿se basan en métodos objetivos o subjetivos? Explique.

Determine las probabilidades de los siguientes eventos:

b. La compañía logra utilidades iguales a las de este año y el sindicato pide aumento de sueldos.

c. La compañía logra mayores utilidades el próximo año y el sindicato no demanda aumento de salarios.

d. El sindicato pide aumento de salarios.

9. Un estudio de mercado llevado a cabo por una empresa comercializadora de aparatos electrodomésticos mostró que de las cuarenta personas entrevistadas, doce tienen secadora de ropa, veinte tienen lavadora de ropa, dieciséis tienen horno de microondas, ocho tienen tanto secadora como horno de microondas y ocho poseen lavadora y microondas. Definiendo:

Evento S = Tener secadora
Evento L = Tener lavadora
Evento M = Tener horno de microondas

Basándose en la información presentada por el estudio de mercado:

a. Encontrar p(S), p(L), p(M), p(SM), p(LM).

b. ¿Son los eventos S y M mutuamente excluyentes?

c. ¿Son los eventos M y L independientes?

d. Si una persona tiene una secadora de ropa, ¿cuál es la probabilidad de que tenga también un horno de microondas?

10. La universidad está planeando adquirir un sistema de emergencia de generación de electricidad que podría entrar en funcionamiento si la electricidad de la ciudad falla. Este sistema tiene dos generadores separados; si uno falla, el otro entra en funcionamiento automáticamente. La probabilidad de que un generador falle es estimada por el fabricante en 0.006. ¿Cuál es la probabilidad de que ambos generadores de emergencia fallen en sucesión, dejando a la universidad sin fluido eléctrico?

11. Se realizó un estudio de mercado con el fin de establecer la relación entre la habilidad de recordar un comercial de televisión para cierto producto y la compra de dicho producto. Se entrevistó a 800 personas, encontrándose los siguientes resultados:

	Pudo recordar el comercial	No pudo recordar el comercial	Total
Compró el producto	160	80	240
No compró el producto	240	320	560
	400	400	800

Si definimos como "R" el evento en que la persona recuerda el comercial y "C" el evento en que la persona compra el producto:

a. Encuentre p(R), p(C) y p(RC).

b. ¿Son los eventos R y C mutuamente excluyentes?

c. ¿Son R y C eventos independientes? Explique utilizando los valores de las probabilidades.

d. ¿Cuál es la probabilidad de que un individuo compre el producto dado que vio el comercial? ¿El que un individuo haya visto el comercial incrementa la probabilidad de que compre el producto?

e. Comente sobre el valor del comercial en relación a las ventas del producto.

12. Simón Pérez es dueño de acciones de la mina "Tantaña" y de la compañía "Año Nuevo". Simón ha mantenido un registro del comportamiento de estas acciones en los últimos 200 días para ver si los precios de sus acciones suben, bajan o se mantienen constantes. Los datos de Simón son los siguientes:

	Mina "Tantaña"	"Año Nuevo"	Nº de días
E1	Subió	Subió	40
E2	Subió	Constante	14
E3	Subió	Bajó	20
E4	Constante	Subió	18
E5	Constante	Constante	12
E6	Constante	Bajó	14
E7	Bajó	Subió	28
E8	Bajó	Constante	16
E9	Bajó	Bajó	38

a. Usando estos datos para determinar la probabilidad subjetiva, encuentre:

- Una subida en el precio de las acciones de "Tantaña".
- Una baja en el precio de las acciones de "Año Nuevo".
- Una subida en las acciones de "Tantaña" y una baja en las de "Año Nuevo".
- Una subida en las acciones de "Tantaña" o una baja en las acciones de "Año Nuevo".

b. Suponga que nos dijeron que las acciones de "Tantaña" han subido. ¿Cuál es la probabilidad de que las de "Año Nuevo" hayan bajado?
c. Presuma que nos dijeron que las acciones de "Año Nuevo" han bajado. ¿Cuál es la probabilidad de que las de "Tantaña" hayan subido?
d. ¿Son los eventos "Una subida en el precio de las acciones de Tantaña" y "Una baja en el precio de las acciones de 'Año Nuevo'", independientes?

13. Margarita tiene una tienda de mascotas y ofrece servicio de lavado y arreglo de perros. A Margarita le toma 40 minutos lavar y arreglar a un perro pequeño y 70 minutos a un perro grande. Los perros grandes representan el 20 % de su negocio. Si atiende a cinco perros en un día determinado, calcular:
a. La probabilidad de que todos sean perros pequeños.
b. La probabilidad de que dos perros sean grandes.
c. El tiempo esperado que demorará Margarita en atender a los cinco perros.

14. Los socios del club "Los Cóndores" de Chaclacayo tienen una probabilidad de 0.80 de gozar de un día de sol, mientras que la probabilidad de lluvia es de 0.10. Además, la probabilidad de tener sol y lluvia en el mismo día es 0.05. Suponga que el experimento a analizar se refiere a las probabilidades climatológicas de un día determinado.
a. ¿Son los eventos sol y lluvia mutuamente excluyentes? Explique.
b. Sabemos que llovió un día determinado. ¿Cuál es la probabilidad de que también haya brillado el sol?
c. ¿Son los eventos sol y lluvia independientes?

15. Considere el experimento de lanzar dos dados. Suponga que estamos interesados en la suma de los números que aparecen en los dados.
a. Enumere los puntos muestrales posibles.
b. ¿Cuál es la probabilidad de obtener un 7?
c. ¿Cuál es la probabilidad de obtener 8 ó más?
d. Dado que hay seis valores pares posibles para la suma de ambos números y sólo cinco valores impares posibles, los dados darán sumas

pares más frecuentemente que sumas impares. ¿Está de acuerdo con esta afirmación? Explique.

16. De todos los postulantes que rinden el examen de ingreso a la Escuela de Postgrado de la Universidad del Pacífico, 3/4 son ingenieros y 1/4 tienen otras profesiones. La mitad de los no-ingenieros y 3/4 de los ingenieros aprueban el examen.

a. Encuentre la probabilidad de ingreso de un postulante escogido aleatoriamente de todos los que rinden el examen.

b. ¿Cuál es la probabilidad de que un estudiante no sea ingeniero y apruebe el examen?

c. ¿Cuál es la probabilidad de que un estudiante que no es ingeniero apruebe el examen?

d. ¿Cuál es la probabilidad de un estudiante que no haya aprobado el examen sea ingeniero?

17. En la última maratón de "Cafetal" el 40% de los participantes no eran limeños. Veinte por ciento de los participantes terminaron la carrera. Ochenta por ciento de los que terminaron la carrera no eran limeños.

a. ¿Cuál es la probabilidad de que un participante elegido aleatoriamente haya terminado la carrera y sea limeño?

b. ¿Cuál es la probabilidad de que un participante elegido aleatoriamente pertenezca por lo menos a una de las siguientes categorías: "terminó la carrera" o "no sea limeño"?

c. ¿Cuál es la probabilidad de que un competidor que no es limeño haya terminado la carrera?

d. ¿Cuál es la probabilidad de que un competidor que es limeño haya terminado la carrera?

e. ¿Cuál es la probabilidad de que un competidor que no haya terminado la carrera sea limeño?

18. En la planta "Cigüeñales Perfectos" se producen cigüeñales en dos grandes máquinas automáticas. La máquina nueva es más rápida y confiable que la vieja. De cada lote de 1,000 cigüeñales, 600 se producen en la máquina nueva. Esta produce cigüeñales defectuosos a razón de 1 por 100, mientras que la máquina vieja lo hace a razón de 3 por 100. ¿Cuál es la probabilidad de que una pieza defectuosa escogida al azar provenga de la máquina nueva?

19. Antes de iniciar el proceso de producción de cada lote, una cierta máquina necesita ser calibrada. La probabilidad de realizar este ajuste correctamente es de 0.90%. Cuando se ha calibrado adecuadamente, la máquina opera con una tasa del 2% de producción defectuosa; pero si está incorrectamente ajustada, se produce un 15% de piezas defectuosas.

a. Después de que la máquina empieza a producir, ¿cuál es la probabilidad de observar un defecto cuando se prueba una pieza?
b. Una pieza seleccionada por un inspector es encontrada defectuosa. ¿Cuál es la probabilidad de que la máquina esté incorrectamente calibrada? ¿Qué acción recomendaría usted?
c. Antes de seguir su recomendación en (b), se prueba una segunda pieza, la que es encontrada sin defecto. Usando su probabilidad revisada de (b) como la probabilidad *a priori* más reciente, calcule una nueva probabilidad revisada de una calibración incorrecta, dado que la segunda pieza que se probó se encontró sin defecto. ¿Qué acción recomendaría ahora?

20. La compañía manufacturera "Los Altos" ha recibido recientemente cinco cajas de una cierta pieza que utiliza en su proceso productivo. La tasa de piezas defectuosas es normalmente del 1%, pero el abastecedor acaba de notificar a "Los Altos" que una de las cajas enviadas contiene piezas que fueron producidas en una máquina mal alineada y que tiene una tasa de piezas defectuosas del 67%. Sabiendo esto, el Gerente de Planta selecciona una caja aleatoriamente y prueba una pieza.
a. ¿Cuál es la probabilidad de que la pieza sea defectuosa?
b. Considere que la pieza resultó defectuosa. ¿Cuál es la probabilidad de que esta sea la caja que contiene las piezas hechas por la máquina mal alineada?
c. Suponga que basado en otras evidencias, el Gerente de Planta está 80% seguro de que esta era la caja que contiene las piezas hechas en la máquina mal alineada. ¿Cómo cambiaría su respuesta a la pregunta (b)?

21. Una firma consultora ha presentado una propuesta para realizar un proyecto importante en Chimbote. El Gerente de la firma pensó que había una probabilidad del 50% de ganar la buena pro. Posteriormente, la agencia a la cual fue presentada la propuesta solicitó información adicional a la propuesta original.

La experiencia de la firma muestra que en el 75% de las propuestas ganadoras y en el 40% de las perdedoras, le fue solicitada información adicional.
a. ¿Cuál es la probabilidad *a priori* de que la propuesta será ganadora? (Es decir, previa a recibir la solicitud de información adicional.)
b. ¿Cuál es la probabilidad condicional de una solicitud de información adicional, dado que la propuesta sea ganadora?
c. Calcule la probabilidad posterior de que la propuesta sea ganadora dado que se recibió una solicitud de información adicional.

22. Una compañía petrolera ha comprado el derecho de explotación en el lote XY de la selva peruana. Los estudios geológicos preliminares han permitido asignar las siguientes probabilidades *a priori*:

p (Petróleo de alta calidad) = 0.50
p (Petróleo de mediana calidad) = 0.20
p (Nada de petróleo) = 0.30

a. ¿Cuál es la probabilidad de encontrar petróleo?
b. Después de perforar 200 pies en el primer pozo, se realiza una prueba de estructuras. Las probabilidades de encontrar una subestructura petrolera son las siguientes:

p (Subestr./petróleo de alta calidad) = 0.20
p (Subestr./petróleo de mediana calidad) = 0.80
p (Subestr./nada de petróleo) = 0.20

¿Cómo debe interpretar la compañía la prueba de estructuras? ¿Cuáles son las probabilidades revisadas y cuál es la nueva probabilidad de encontrar petróleo?

23. Un asesor de inversiones reputado por su vasto conocimiento sobre el comportamiento de las acciones de las minas "Macato y Cía." nos da la siguiente información:

p (Las acciones suben 20%/PNB sube) = 0.6
p (Las acciones suben 20%/PNB estable) = 0.5
p (Las acciones suben 20%/PNB baja) = 0.4

Un distinguido economista que estudia las variables macroeconómicas establece que la probabilidad de un aumento en el PNB es del 0.30%, mientras que la probabilidad de que baje es de 0.40%.
a. ¿Cuál es la probabilidad de que la cotización de las acciones de las minas suba 20%?
b. Si nos dijeron que las cotizaciones de las acciones subieron el 20%, ¿cuál es la probabilidad de un aumento o disminución del PNB?

24. Una compañía de auditoría se ha percatado de que el 85% de las empresas que audita no tiene déficit de inventarios, mientras que el 10% tiene déficit pequeños y el 5% tiene déficit grandes. La compañía de auditoría ha desarrollado una nueva prueba de contabilidad para la cual se cree que las siguientes probabilidades existen:

p (Empresa pasará la prueba/sin déficit) = 0.90
p (Empresa pasará la prueba/déficit pequeño) = 0.50
p (Empresa pasará la prueba/déficit grande) = 0.20

a. Si la empresa que se está auditando no pasa esta prueba, ¿cuál es la probabilidad de que tenga un déficit grande o pequeño de inventarios?
b. Si la empresa que se está auditando pasa esta prueba, ¿cuál es la probabilidad de que no tenga déficit de inventarios?

25. Suponga que el 80% de los clientes de "SAGA" son dignos de crédito. Suponga además que la probabilidad de que un comprador digno de crédito tenga cuenta bancaria es de 75%, mientras que esta probabilidad es de 35% para los clientes no dignos de crédito.
a. Construya un árbol de probabilidades para esta situación.
b. ¿Cuál es la probabilidad de que un cliente elegido al azar tenga cuenta bancaria?
c. Calcule la probabilidad de que un cliente con cuenta bancaria no sea digno de crédito.
d. Calcule la probabilidad de que un solicitante de crédito que no tiene cuenta bancaria sea digno de crédito.

26. La compañía de seguros "Segurísimo" ha recolectado las siguientes estadísticas para un período cualquiera de un año:

p (Accidente/conductor hombre menor de 25 años) = 0.22
p (Accidente/conductor hombre mayor de 25 años) = 0.15
p (Accidente/conductora mujer menor de 25 años) = 0.16
p (Accidente/conductora mujer mayor de 25 años) = 0.14

El porcentaje de asegurados de "Segurísimo" en cada categoría es el siguiente:

Hombres mayores de 25 : 40%
Hombres menores de 25 : 20%
Mujeres mayores de 25 : 30%
Mujeres menores de 25 : 10%

a. ¿Cuál es la probabilidad de que una persona asegurada escogida aleatoriamente tenga un accidente en el próximo año?
b. Dado que una persona asegurada tiene un accidente, ¿cuál es la probabilidad de que sea un hombre menor de 25 años?
c. Dado que una persona asegurada no tiene ningún accidente, ¿cuál es la probabilidad de que sea mujer?

d. Sabiendo que una persona asegurada no ha tenido ningún accidente, ¿nos ofrece suficiente información respecto a su sexo?

27. A es culpable de un crimen con una probabilidad *a priori* de 0.40. B y C, quienes saben si A es culpable, han sido llamados a testificar. B es amigo de A y dirá la verdad si A es inocente, pero mentirá con una probabilidad de 0.30 si A es culpable. C odia a todos menos al juez y dirá la verdad si A es culpable pero mentirá con una probabilidad de 0.40 si A es inocente. B y C testifican independientemente.

a. ¿Cuál es la probabilidad de que ambos testigos mientan?

b. Dado el hecho de que B y C han presentado testimonios contradictorios:

- ¿Cuál es la probabilidad de que B esté diciendo la verdad?
- ¿Cuál es la probabilidad de que A sea inocente?

III. Variables aleatorias

1. El espacio muestral. 2. El árbol de probabilidades. 3. Variables aleatorias. 4. Representación de las variables aleatorias discretas. 5. La distribución binomial. 6. La distribución de Poisson. 7. Variables aleatorias continuas: La distribución normal.

En este capítulo continuamos con el estudio de probabilidades como la herramienta básica para incorporar el aspecto de incertidumbre del medio ambiente en el análisis del proceso de toma de decisiones. En el capítulo anterior precisamos que un experimento era cualquier proceso que genera resultados bien definidos. Ahora nos concentraremos en los procedimientos para asignar valores numéricos a dichos resultados. Es así como surge el concepto de variable aleatoria.

Las variables aleatorias, tanto discretas como continuas, son útiles para la construcción de modelos de sistemas específicos. Por ejemplo, un sistema de inventarios involucra variables cuyos valores no podemos conocer con certeza, como el nivel de demanda, costos, etcétera.

Se introduce el concepto de espacio muestral para formar la base de la valoración probabilística y del análisis. El espacio muestral puede ser usado para reducir la solución de problemas complejos a una secuencia de pasos simples, como se muestra en la sección 1. Muchas veces es conveniente representar el

espacio muestral en la forma de un árbol de probabilidades, el cual se describe en la sección 2. En la sección 3 se introduce el concepto de variable aleatoria discreta. Este tipo de variables son caracterizadas por medio de la esperanza matemática, la varianza y las funciones de probabilidades en la sección 4. Dos distribuciones de probabilidades discretas de uso amplio son presentadas en las secciones 5 y 6: la distribución binomial y la distribución de Poisson.

Muchos modelos se formulan tomando como base variables aleatorias continuas. Prácticamente toda la discusión de variables aleatorias discretas tiene su análoga en el caso continuo, pero debemos usar el cálculo en lugar de las matemáticas discretas. Finalmente se presenta la distribución normal, la más importante de las distribuciones continuas dado que es la base de la estadística inferencial y del análisis de regresión.

1. EL ESPACIO MUESTRAL

La noción del espacio muestral proporciona un concepto fundamental con el cual se pueden resolver problemas tanto teóricos como prácticos. Por ser una noción simple, muchas veces el estudiante de probabilidades no le presta la atención que se merece.

Ampliaremos primero el concepto de *experimento* en estadística, que se presentó en el capítulo II. Todo experimento tiene dos propiedades fundamentales: a) genera dos o más resultados posibles, que pueden ser especificados de antemano; y, b) la ocurrencia de los resultados es incierta. Ejemplos de experimentos estadísticos son el lanzamiento de un dado o una moneda, la selección de una bujía de un lote para determinar si está defectuosa o no, la introducción de un nuevo producto en el mercado. Todo experimento genera un espacio que se construye:

1. Creando un punto muestral para cada resultado posible. Cada resultado es asociado con uno y sólo un punto muestral,

y cada punto muestral corresponde a un resultado específico. Es decir, los puntos muestrales representan eventos mutuamente excluyentes y colectivamente exhaustivos.

2. Asignando una probabilidad a cada punto muestral. Esta probabilidad es la que corresponde al resultado del experimento. Si lanzamos dos monedas simultáneamente, tendremos cuatro resultados posibles y, por lo tanto, cuatro puntos muestrales: que ambas muestren cara (CC), que ambas muestren sello (SS), que una muestre cara y la otra sello (CS) y viceversa (SC). Luego debemos asignar probabilidades a la ocurrencia de cada resultado. Si suponemos que las monedas están balanceadas, asignaremos una probabilidad de 0.25 a cada uno de los puntos en el espacio muestral.

Una vez construido el espacio muestral, este se podrá utilizar para calcular la probabilidad de *eventos* o *sucesos* asociados con el experimento. Podemos definir un suceso como el conjunto de resultados del experimento que tengan una propiedad particular. Luego, para encontrar la probabilidad de este suceso se deberá examinar cada punto muestral para determinar si tiene la propiedad deseada y por lo tanto origina que el suceso ocurra. De esta manera se establece el conjunto de puntos muestrales correspondientes a la ocurrencia del suceso. La probabilidad del suceso es entonces calculada como la suma de las probabilidades asociadas con cada uno de estos puntos muestrales, dado que se trata de eventos mutuamente excluyentes.

En resumen, el espacio muestral permite calcular la probabilidad de cualquier suceso asociado con un experimento, lo que se consigue: a) representando el suceso como una colección de puntos en el espacio muestral; y, b) determinando la probabilidad del suceso como la suma de las probabilidades de los respectivos puntos muestrales.

El concepto de espacio muestral puede ilustrarse mejor con un ejemplo. Supongamos que el experimento consiste en lanzar dos dados no cargados, uno rojo y otro verde. El espacio mues-

tral para este experimento aparece en el cuadro 4. En dicho cuadro vemos que existen 36 puntos muestrales. Puesto que son dados no cargados y que los números que se obtienen en los dos dados son independientes entre sí, asignaremos la probabilidad de 1/36 a cada punto muestral que se aprecia en el cuadro 4.

CUADRO 4: ESPACIO MUESTRAL PARA EL LANZAMIENTO DE DOS DADOS NO CARGADOS.
(Cada punto muestral tiene una probabilidad de 1/36)

Dado rojo	Dado verde					
	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Ahora estamos listos para encontrar la probabilidad de cualquier *suceso* relacionado con los resultados de este experimento. Por ejemplo, consideremos el suceso S_1, definido como aquel en el que la suma de los puntajes obtenidos en los dados es 7. Para calcular la probabilidad de este suceso debemos: (i) encontrar los puntos muestrales asociados con este suceso; y (ii) sumar las probabilidades de estos puntos:

(i) $S_1 = (1,6) + (2,5) + (3,4) + (4,3) + (5,2) + (6,1)$

(ii) $p(S_1/e) = 1/36 + 1/36 + 1/36 + 1/36 + 1/36 + 1/36 = 1/6$

Si lanzamos dos dados no cargados, la probabilidad de que la suma de los puntajes de ambos dados sea 7 es 1/6.

Definamos el suceso S_2 como aquel en el que la suma de los puntajes en los dos dados sea divisible por 4. Hay 9 puntos muestrales asociados a S_2, cada uno con una probabilidad de 1/36.

(i) $S_2 = (1,3) + (2,2) + (2,6) + (3,1) + (4,4) + (3,5) + (5,3) + (6,2) + (6,6)$

(ii) $p(S_2/e) = 9(1/36) = 1/4$

Supongamos ahora que los dados rojo y verde estén cargados a favor de los números grandes. Supongamos también que las probabilidades de obtener un cierto número al lanzar los dados están en función de ese número (i) y pueden definirse con la siguiente fórmula:

$$p(R_i/e) = p(V_i/e) = \frac{2i - 1}{36} \qquad (1)$$

R_i = Puntaje en el dado rojo igual a i, i=1,2,3,4,5,6.
V_i = Puntaje en el dado verde igual a i, i=1,2,3,4,5,6.

Entonces las probabilidades para cada valor de i son:

$p(1/e) = 1/36$;	$p(2/e) = 3/36$;	$p(3/e) = 5/36$
$p(4/e) = 7/36$;	$p(5/e) = 9/36$;	$p(6/e) = 11/36$

Los puntos muestrales del experimento de lanzar dos dados son todavía los mismos, pero la probabilidad asociada a cada uno de estos puntos ha cambiado, debido a que los dados están cargados. El cuadro 5 muestra las nuevas probabilidades de cada punto muestral.

CUADRO 5: PROBABILIDAD DE CADA PUNTO MUESTRAL AL LANZAR DOS DADOS CARGADOS*

Dado rojo	Dado verde					
	1	2	3	4	5	6
1	1/D	3/D	5/D	7/D	9/D	11/D
2	3/D	9/D	15/D	21/D	27/D	33/D
3	5/D	15/D	25/D	35/D	45/D	55/D
4	7/D	21/D	35/D	49/D	63/D	77/D
5	9/D	27/D	45/D	63/D	81/D	99/D
6	11/D	33/D	55/D	77/D	99/D	121/D

* El valor del denominador común, D, es igual a 1,296.

Retomemos los sucesos S_1 y S_2 definidos anteriormente. Los puntos muestrales asociados a S_1 y sus probabilidades están dados por:

(i) $S_1 = (1,6) + (2,5) + (3,4) + (4,3) + (5,2) + (6,1)$

(ii) $p(S_1/e) = (11 + 27 + 35 + 35 + 27 + 11)/1{,}296$
$= 146/1{,}296$

La probabilidad de que ocurra el suceso S_1, es decir que la suma de los puntajes en los dos dados sea 7 con los dados cargados, es menor que la probabilidad del mismo suceso con los dados no cargados ($216/1{,}296 < 1/6$).

De igual manera, para encontrar la probabilidad de S_2 se requiere sumar las nuevas probabilidades de los 9 puntos muestrales asociados a este suceso:

(i) $S_2 = (1,3) + (2,2) + (2,6) + (3,1) + (3,5) + (4,4) +$
$(5,3) + (6,2) + (6,6)$

(ii) $p(S_2/e) = (3 + 9 + 33 + 15 + 45 + 49 + 45 + 33 + 121)/1{,}296$
$= 353/1{,}296.$

La probabilidad de que la suma de los puntajes obtenidos en cada dado sea divisible por 4 es mayor con los dados cargados que con los dados no cargados ($353/1{,}296 > 1/4$)

2. EL ÁRBOL DE PROBABILIDADES

Una manera conveniente de representar el espacio muestral es mediante un árbol de probabilidades. Puede construirse un árbol de probabilidades para cualquier tipo de experimento, aunque su interpretación es más directa cuando los resultados de las distintas partes del experimento se cumplen secuencialmente. Sin embargo, un experimento cuyas partes ocurren simultáneamente puede ser considerado como si estas ocurrieran secuencialmente. Por ejemplo, en nuestro experimento de lanzar dos dados podemos afirmar arbitrariamente que el número en el dado rojo es revelado antes que el número en el dado verde. Esta convención no afectará ninguna de las probabilidades de los sucesos basados en el experimento original.

El gráfico 20 muestra la forma general de un árbol de probabilidades, aunque ha sido dibujado para el experimento específico de lanzar dos dados. Un árbol de probabilidades se construye de izquierda a derecha, desde un punto llamado *nodo* o *nudo,* del cual se dibuja una línea o rama por cada resultado posible de la primera parte del experimento. Cada rama es denotada con el resultado y con la probabilidad de dicho resultado. Luego se usan los terminales de las primeras ramas como los nodos de comienzo para otros grupos de ramas que representan a cada posible resultado de la segunda parte del experimento. Estas ramas son denotadas con dichos resultados y con las probabilidades *condicionales* de cada segundo resultado dado que ocurrió el primer resultado. El proceso continúa hasta que

todas las secuencias posibles de resultados del experimento hayan sido consideradas. Las probabilidades en las ramas que salen de un nodo son condicionales a los resultados que preceden a dicho nodo. Los nodos terminales del árbol representan todos los puntos muestrales del experimento. La probabilidad de cada nodo terminal se calcula multiplicando todas las probabilidades en el conjunto de ramas que nos llevan desde el nodo inicial hasta el nodo final, aplicando el concepto de expansión en cadena.

El árbol de probabilidades para el lanzamiento de los dos dados de colores tiene 36 nodos terminales que corresponden a los 36 puntos muestrales en el cuadro 4. La probabilidad de los puntos muestrales para el caso de los dados no cargados es calculada con base en:

$$\begin{aligned} p(R_i/e) &= 1/6 \\ p(V_j/R_i,e) &= p(V_j/e) = 1/6 \end{aligned} \qquad (3)$$

Entonces:

$$p(R_iV_j/e) = p(R_i/e)\, p(V_j/e) = 1/6 * 1/6 = 1/36$$

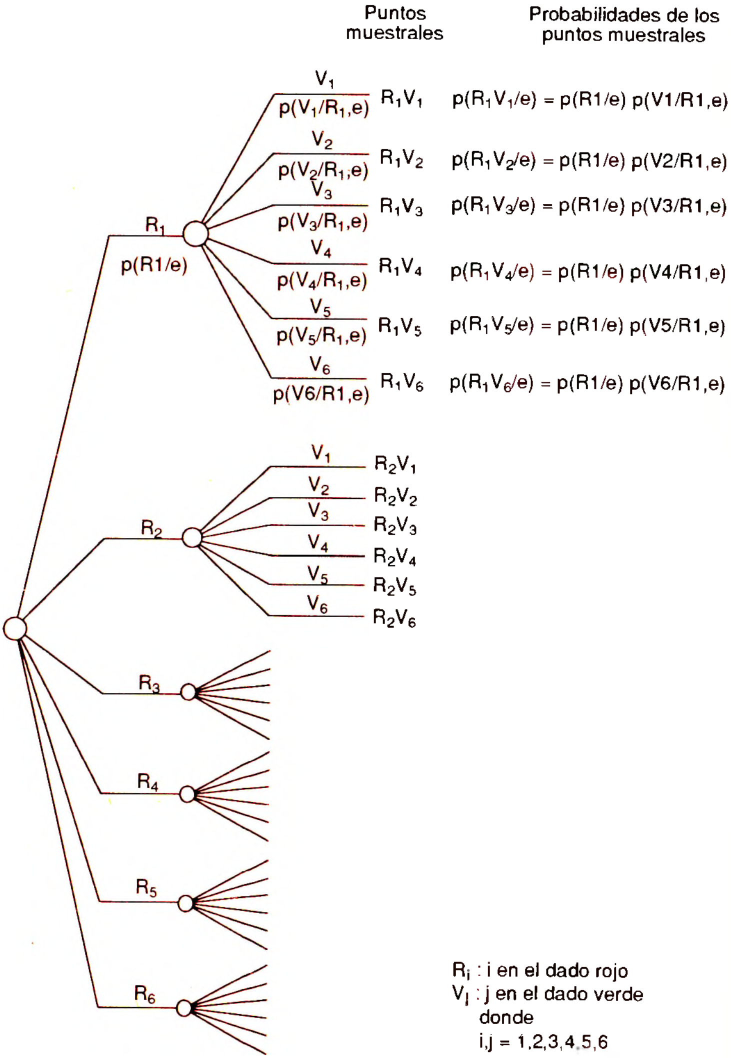
La probabilidad de cada nodo final será de 1/36, como se muestra en el cuadro 4.

Para el caso de los dados cargados:

$$\begin{aligned} p(R_i/e) &= p(i/e) \\ p(V_j/R_i,e) &= p(V_j/e) = p(j/e) \end{aligned} \qquad (4)$$

donde p(1/e), p(2/e), ..., p(6/e) están definidos por la ecuación (1). Al calcular las probabilidades de los nodos finales del árbol de probabilidades del gráfico 20 obtendremos los mismos valores que se presentan en el cuadro 5.

Gráfico 20: Árbol de probabilidades para la tirada de dos dados



La utilización del árbol para el cálculo de la probabilidad de ocurrencia de cualquier suceso asociado a los resultados del experimento se realiza de igual manera a como se usó con el espacio muestral en su forma regular. Para obtener la probabilidad de ocurrencia del suceso se suman las probabilidades de los puntos muestrales o nodos finales correspondientes a tal suceso. La ventaja de utilizar un árbol de probabilidades es que permite representar en forma más clara y conveniente problemas multidimensionales; es decir, experimentos con más de dos partes.

Para una mejor ilustración del uso del árbol de probabilidades, considérese el siguiente ejemplo: Se tiene una caja con cuatro medias negras (N), seis medias blancas (B) y dos medias rojas (R). El experimento consiste en extraer dos medias aleatoriamente, secuencialmente y sin reemplazo. El gráfico 21 muestra el árbol de probabilidades para este ejemplo, que consta de ocho nodos finales que representan los puntos muestrales del experimento. La probabilidad de cada punto muestral se calcula multiplicando las probabilidades desde el nodo inicial hasta el nodo final respectivo. Por ejemplo, la probabilidad de obtener dos medias rojas, R_1R_2, está dada por:

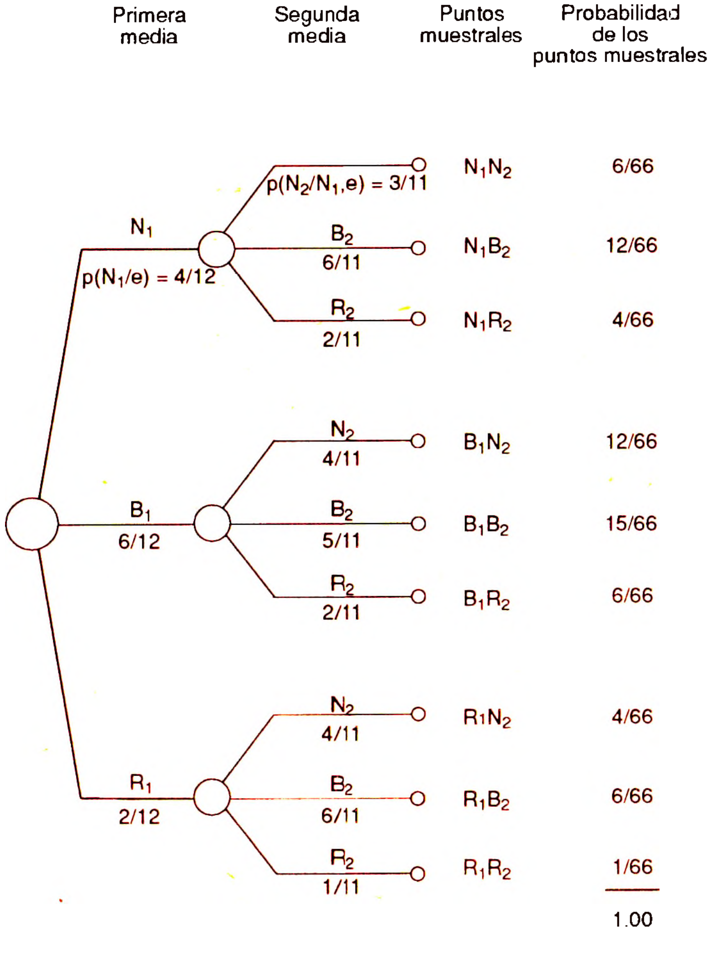
$$p(R_1, R_2/e) = p(R_1/e)\ p(R_1/R_2,e)$$

Con el árbol de probabilidades podemos calcular la probabilidad de ocurrencia de cualquier suceso asociado con este experimento. Suponga que deseamos calcular la probabilidad de que ambas medias sean del mismo color. Denotando a este suceso S_3 tenemos:

$$S_3 = (N_1, N_2) + (B_1, B_2) + (R_1, R_2)$$

$$\begin{aligned} p(S_3/e) &= p(N_1\ N_2/e) + p(B_1\ B_2/e) + p(R_1\ R_2/e) \\ &= 6/66 + 15/66 + 1/66 = 22/66 = 1/3 \end{aligned}$$

Gráfico 21: Árbol de probabilidades para sacar dos medias de una caja



La probabilidad de obtener ambas medias del mismo color es 1/3.

Consideremos otro suceso como S_4 que consiste en obtener una media negra y una blanca. La probabilidad de S_4 será:

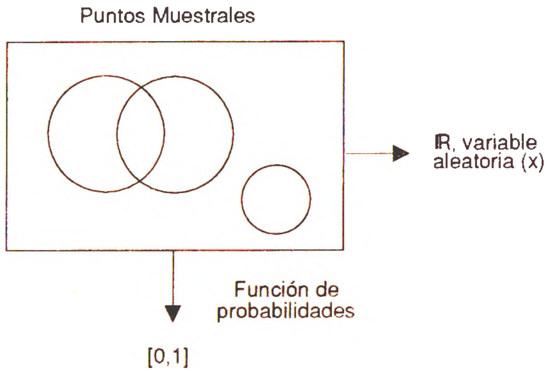
$$S_4 = (N1, B_2) + (B_1, N_2)$$

$$p(S_4/e) = p(N_1 B_2/e) + p(B_1 N_2/e) = 12/66 + 12/66 = 4/11$$

3. VARIABLES ALEATORIAS

Una variable aleatoria es una variable cuyo comportamiento es descrito por una función de probabilidades, y puede ser definida asignando algún número real a cada punto del espacio muestral. Entonces, a cada punto muestral se le asigna una probabilidad y un número real. Estos números reales constituyen los posibles valores de la llamada variable aleatoria. Esquemáticamente:

Gráfico 22: La función de probabilidad y la variable aleatoria



Es posible describir completamente un espacio muestral por medio de variables aleatorias. La especificación del espacio muestral y de las variables aleatorias depende de los sucesos que son considerados.

Podemos ilustrar el concepto de variable aleatoria usando nuestro experimento de los dos dados. Sea X la variable aleatoria que representa el producto de los puntajes en los dos dados. El valor de esta variable para cada punto muestral se muestra en el cuadro 6, y en la parte derecha del gráfico 23.

CUADRO 6: UNA VARIABLE ALEATORIA PARA EL EXPERIMENTO DE DOS DADOS
(Los valores en cada celda representan el valor de la variable aleatoria X = producto de los puntajes de los dados en el punto muestral)

Puntaje en el dado rojo	Puntaje en el dado verde					
	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	10	12
3	3	6	9	12	15	18
4	4	8	12	16	20	24
5	5	10	15	20	25	30
6	6	12	18	24	30	36

El valor de la variable aleatoria X, producida por el experimento, varía entre 1 y 36, pero no todos los números enteros entre estos límites son valores válidos para la variable X. Notamos que existen valores de la variable aleatoria que pueden ser asociados con varios puntos muestrales. Sin embargo, cada punto muestral puede asociarse con un solo valor.

Consideremos el experimento de extraer dos medias de una caja. Sea Y la variable aleatoria que representa el número de medias negras obtenidas, y Z el número de pares de medias, es

Gráfico 23: Variables aleatorias asociadas al experimento de lanzar dos dados

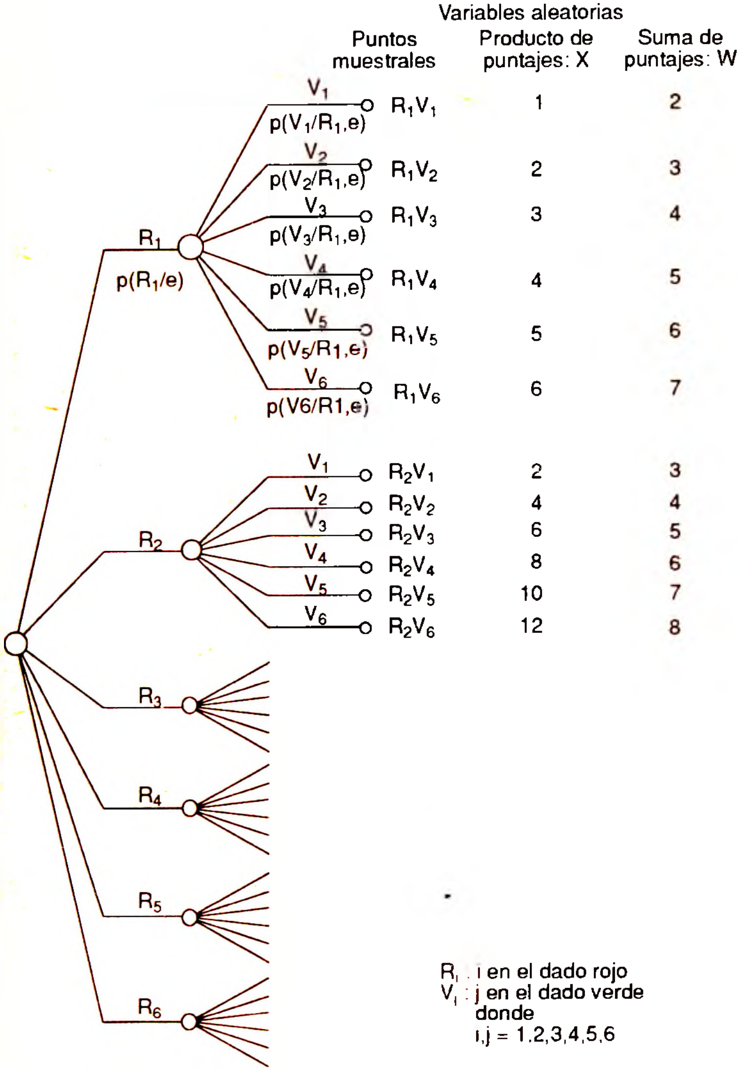
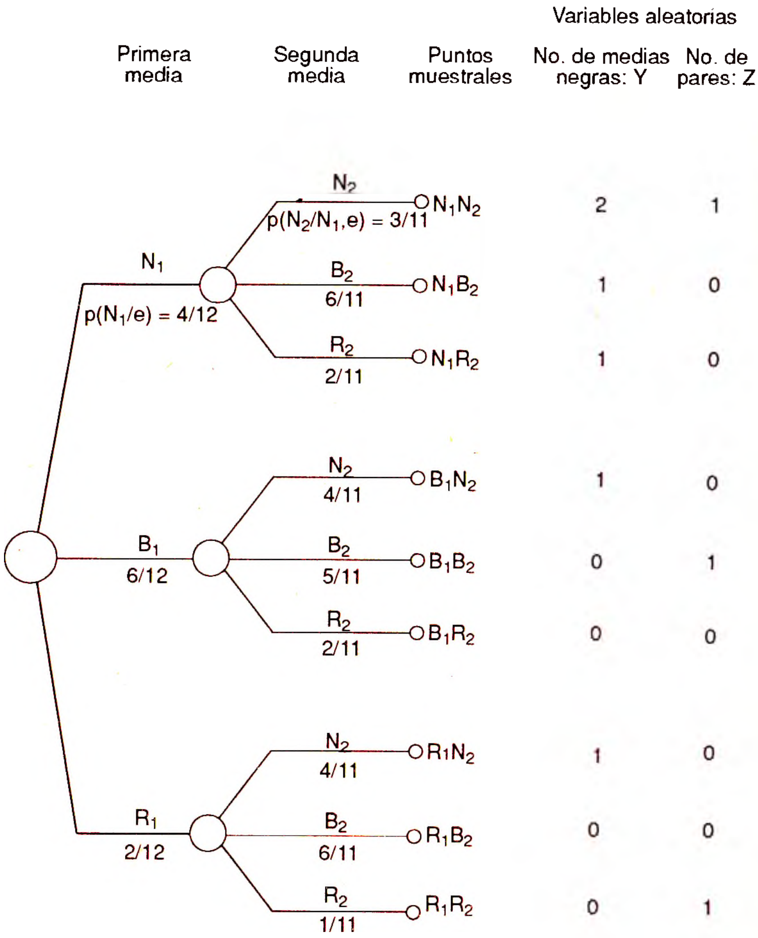


Gráfico 24: Variables aleatorias asociadas con el experimento de sacar dos medias de una caja



decir, obtener ambas medias del mismo color. El valor de estas variables para cada punto muestral se presenta en la parte derecha del gráfico 25. La variable Y puede tomar los valores 0, 1 y 2, mientras que la variable Z sólo toma los valores 0 y 1. Una vez definida una variable aleatoria, el siguiente paso consiste en describirla en forma sucinta usando medidas de tendencia central y dispersión.

Esperanza matemática

La esperanza matemática de una variable es un concepto simple pero de gran importancia en la teoría de probabilidades. Se define como la suma de los productos de la probabilidad por el valor de la variable aleatoria asociada con cada punto del espacio muestral. Si M_i es el i-ésimo punto muestral, $X(M_i)$ el valor de la variable aleatoria en ese punto y $p(M_i/e)$ la probabilidad del punto muestral, entonces la esperanza matemática de la variable aleatoria X es definida como:

$$E(X) = \sum X(M_i)\, p(M_i/e) \qquad (5)$$

La esperanza matemática de X es también conocida como el *valor esperado* de X, la *media* de X y el *valor promedio* de X. Simbólicamente, se representa de varias formas:

$$E(X) = \bar{X} = \langle X \rangle \qquad (6)$$

El valor de la esperanza matemática de una variable aleatoria en un experimento que se repite muchas veces representa el valor promedio de los valores producidos por dicho experimento. Es decir, si sumamos los valores de la variable aleatoria resultante de estos experimentos y los dividimos entre el número de experimentos realizados, esperamos que el valor de esta cantidad se aproxime cada vez más a la esperanza matemática

de la variable aleatoria, a medida que el número de experimentos aumenta.

En el experimento de los dos dados la esperanza matemática de la variable aleatoria X, definida como el producto de los puntajes en los dados, dependerá de si los dados están cargados o no. Si se trata de dados no cargados, el valor esperado es calculado multiplicando cada uno de los valores de X en el gráfico 23 por la probabilidad de cada punto muestral, 1/36. Así:

$$
\begin{aligned}
E(X) &= 1/36\,(1 + 2 + 3 + 4 + 5 + 6 + 2 + 4 \ldots + 30 + 36) \\
&= 1/36\,(441) = 12.25
\end{aligned}
$$

Si los dados están cargados, entonces el valor esperado de X debe ser calculado sumando el producto de cada valor de la variable aleatoria del gráfico 23 por la probabilidad correspondiente del cuadro 5:

$$
\begin{aligned}
E(X) &= 1\,(1/1{,}296) + 2\,(3/1{,}296) + 3\,(5/1{,}296) + \ldots \\
&= 20.00
\end{aligned}
$$

Puesto que los dados cargados están sesgados hacia los números mayores, producen un valor esperado de X, mayor que el calculado cuando el experimento se realiza con dados no cargados.

También podemos calcular el valor esperado para las variables asociadas con el experimento de extraer dos medias de una caja. Los nueve resultados o puntos muestrales de este experimento, sus probabilidades y los valores de las variables Y y Z asociados con dichos resultados se muestran en el cuadro 7.

Tomando como base estos datos podemos escribir en forma sucinta los valores de las variables y las probabilidades correspondientes a cada uno de estos valores, sumando las probabilidades de los puntos muestrales que generan un mismo valor para la variable aleatoria, como se muestra en el cuadro 8.

CUADRO 7: PUNTOS MUESTRALES Y VARIABLES ALEATORIAS DEL EXPERIMENTO DE EXTRAER DOS MEDIAS DE UNA CAJA

Puntos muestrales	Probabilidades de los puntos	Variables aleatorias	
		Nº de medias negras: Y	Nº de pares: Z
$N_1 N_2$	6/66	2	1
$N_1 B_2$	12/66	1	0
$N_1 R_2$	4/66	1	0
$B_1 N_2$	12/66	1	0
$B_1 B_2$	15/66	0	1
$B_1 R_2$	6/66	0	0
$R_1 N_2$	4/66	1	0
$R_1 B_2$	6/66	0	0
$R_1 R_2$	1/66	0	1

CUADRO 8: DISTRIBUCIÓN DE PROBABILIDADES PARA EL NÚMERO DE MEDIAS NEGRAS (Y) Y EL NÚMERO DE PARES (Z), OBTENIDAS AL EXTRAER DOS MEDIAS DE UNA CAJA

Y: Nº de medias negras	p(Y/e)	Z: Nº de pares	p(Z/e)
0	28/66	0	44/66
1	32/66	1	22/66
2	6/66		
Total	1		1

Una vez que tenemos todos los valores posibles de la variable aleatoria con sus respectivas probabilidades, podemos calcular más fácilmente la esperanza matemática sumando los productos de cada valor posible de dicha variable por su probabilidad de ocurrencia. Así:

$$E(Y) = \sum Y\, p(Y/e) = 0\,(28/66) + 1\,(32/66) + 2\,(6/66)$$
$$= 44/66 = 2/3 \text{ medias negras}$$

$$E(Z) = \sum Z\, p(Z/e) = 0\,(44/66) + 1\,(22/66)$$
$$= 1/3 \text{ pares}$$

Estos valores son idénticos a los que se obtendrían si usamos la ecuación (5).

4. REPRESENTACIÓN DE LAS VARIABLES ALEATORIAS DISCRETAS

Si un espacio muestral tiene un número finito de puntos muestrales, entonces cualquier variable asociada con dicho espacio muestral sólo puede tomar un número finito de valores. A dicha variable se le llama variable aleatoria discreta. Hay varias maneras de representar este tipo de variables. Como ya vimos, una de ellas es asignar un valor de la variable a cada punto muestral. Si todas las preguntas con respecto al experimento pueden ser expresadas en términos de los valores de la variable aleatoria, entonces la representación del espacio muestral a través de esta será conveniente y suficiente. En el caso del experimento de extraer dos medias de una caja, se definieron las variables Y y Z, las cuales nos permiten contestar preguntas con respecto al número de medias negras y el número de pares que se obtienen.

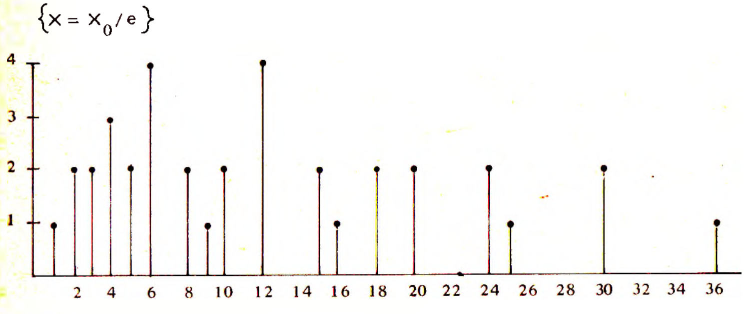
Las representaciones alternativas disponibles de la variable aleatoria son las *distribuciones de probabilidades*. En el experimento de los dados, notamos que a pesar de que hay 36 puntos muestrales distintos, en el cuadro 6 se observa que existen menos de 36 valores de la variable aleatoria X. Si representamos estos valores en el eje de las abscisas y sus respectivas probabilidades en el eje de las ordenadas, obtenemos un diagrama como el del gráfico 25 para el caso de los dados no cargados. Este diagrama es conocido como la función de distribución de pro-

babilidades de la variable aleatoria, porque muestra cómo se distribuye la probabilidad total, que es igual a uno, entre los diferentes valores posibles de la variable aleatoria. Recordemos que la función de probabilidades se calcula sumando las probabilidades de los puntos muestrales asociados con cada valor de la variable aleatoria. Denotamos $p(X = X_0/e)$ como la probabilidad de que la variable aleatoria X tome el valor X_0.

Cualquier pregunta acerca de la probabilidad de obtener un determinado producto de los puntajes de los dados puede ser contestada conociendo la función de distribución de probabilidades de esta variable aleatoria. Por ejemplo, la probabilidad de que el producto de los puntajes de los dados se encuentre entre 22 y 32, inclusive, es la suma de las alturas de los segmentos verticales entre dichos puntos en la función de distribución de probabilidades del gráfico 25.

$$\begin{aligned} p(22 \leq X \leq 32/e) &= p(X = 24/e) + p(X = 25/e) + p(X = 30/e) \\ &= 2/36 + 1/36 + 2/36 = 5/36 \qquad (7) \end{aligned}$$

Gráfico 25: Función de la distribución de probabilidades de la variable aleatoria x, el producto de los puntajes de dos dados no cargados



Sin embargo, si queremos hacer preguntas acerca de las probabilidades de obtener diferentes valores para la *suma* de los

puntajes en los dados, tendremos que regresar al espacio muestral original o construir una función de distribución de probabilidades para esta nueva variable.

A. Distribución de probabilidades acumuladas

La función de distribución de probabilidades no es la única representación útil de una variable aleatoria discreta. Cualquier variable aleatoria puede ser representada efectivamente a través de una distribución de probabilidades acumulada. Denotemos con $p(X \leq X_0/e)$ la probabilidad de que la variable aleatoria X tome valores menores o iguales a X_0. La función $p(X \leq X_0/e)$ es conocida como la *distribución de probabilidades acumuladas* de la variable aleatoria. Para una variable aleatoria discreta, la distribución de probabilidades acumulada se relaciona a la función de distribución de probabilidad de la siguiente manera:

$$p(X \leq X_0/e) = \sum_{X_i \leq X_0} p(X = X_i/e) \qquad (8)$$

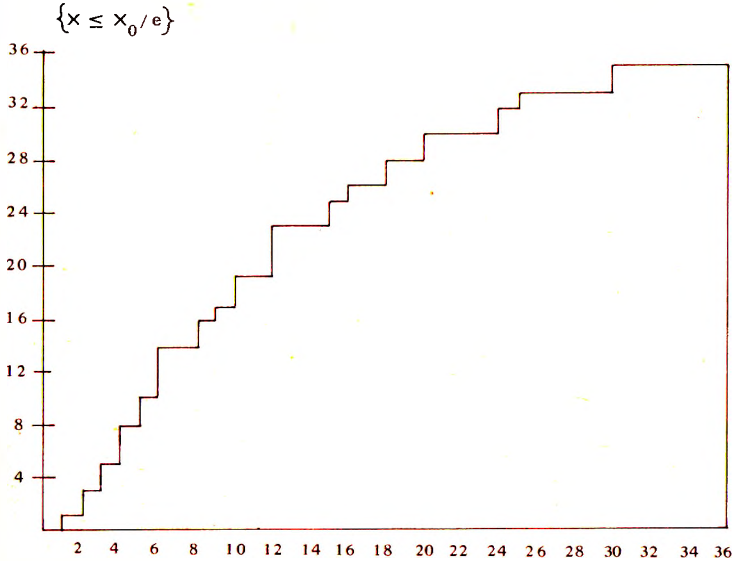
Si X_0 está entre dos enteros, tomamos el valor de la distribución acumulada al valor del entero menor. Con esta definición, $p(X \leq X_0/e)$ es simplemente la suma de las alturas de los segmentos de la función de distribución de probabilidades desde el comienzo del diagrama hasta el valor de X_0. El gráfico 26 muestra la distribución de probabilidad acumulada para el producto de los puntajes de dos dados no cargados calculada usando la función de distribución de probabilidades del gráfico 25. La altura de cada peldaño en la distribución de probabilidades acumulada es igual a la altura del segmento en los puntos correspondientes de la función de distribución de probabilidades $p(X = X_0/e)$.

La probabilidad de que el producto de los puntajes obtenidos al lanzar dos dados no cargados esté entre 22 y 32, inclusive,

puede ser calculada usando la distribución de probabilidades acumulada del gráfico 26:

$$\begin{aligned} p(22 \leq X \leq 32/e) &= p(X \leq 32/e) - p(X \leq 21/e) \\ &= 35/36 - 30/36 \\ &= 5/36 \end{aligned}$$

Gráfico 26: Distribución de probabilidades acumulada de la variable aleatoria x, el producto de los puntajes de dos dados no cargados



Este resultado es igual al obtenido en la ecuación (7).

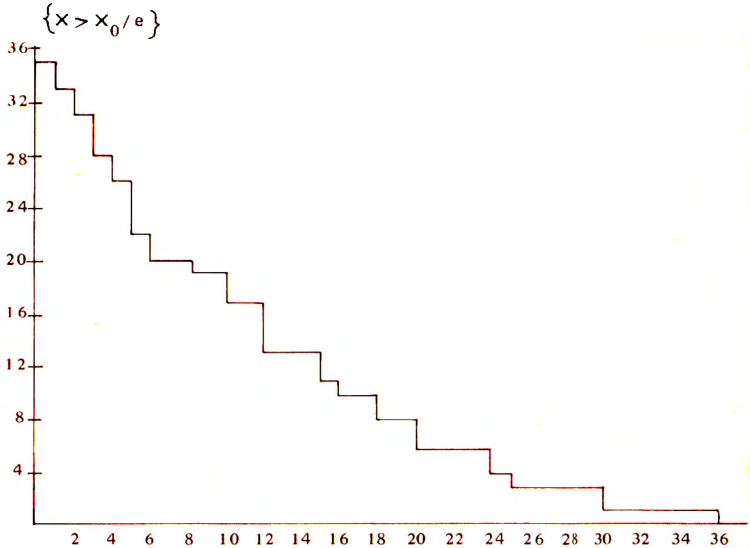
Consideremos un problema en el que sea necesario hallar la probabilidad de que una variable aleatoria exceda un valor dado. Denotemos $p(X > X_0/e)$ como la probabilidad de que la variable aleatoria X sea estrictamente mayor que el número X_0. La función $p(X > X_0/e)$ es conocida como la *distribución de*

probabilidad acumulada complementaria de X; su valor en cualquier punto es igual a 1 menos el valor de la distribución de probabilidad acumulada en ese punto:

$$p(X > X_o/e) = 1 - p(X \leq X_o/e) \qquad (9)$$

La distribución de probabilidades acumuladas complementaria de X calculada usando esta ecuación es mostrada en el gráfico 27.

Gráfico 27: Distribución de probabilidades acumulada complementaria de x, el producto de los puntajes de dos dados no cargados



Los tres tipos de funciones de distribuciones de la variable X, que representa el producto de los puntajes obtenidos, se muestran en el cuadro 9.

CUADRO 9: DISTRIBUCIONES DE PROBABILIDAD DE X, EL PRODUCTO DE LOS PUNTAJES DE DOS DADOS NO CARGADOS

X	$p(X = X_o/e)$	$p(X \leq X_o/e)$	$p(X > X_o/e)$
1	1/36	1/36	35/36
2	2/36	3/36	33/36
3	2/36	5/36	31/36
4	3/36	8/36	28/36
5	2/36	10/36	26/36
6	4/36	14/36	22/36
8	2/36	16/36	20/36
9	1/36	17/36	19/36
10	2/36	19/36	17/36
12	4/36	23/36	13/36
15	2/36	25/36	11/36
16	1/36	26/36	10/36
18	2/36	28/36	8/36
20	2/36	30/36	6/36
24	2/36	32/36	4/36
25	1/36	33/36	3/36
30	2/36	35/36	1/36
36	1/36	36/36	0/36

B. Varianza

Mientras que la esperanza matemática es una medida del promedio o valor central de una variable aleatoria, la varianza nos permite tener una idea de la dispersión o variabilidad de los posibles valores de la variable aleatoria alrededor de la esperanza matemática. La varianza se denota como Var (X) o σ^2, y se calcula usando la siguiente expresión:

$$\text{Var}(X) = \sigma^2 = \sum [X_o - E(X)]^2 \; p(X = X_o/e) \qquad (10)$$

Esta ecuación muestra que la varianza es un promedio ponderado de los cuadrados de las desviaciones de cada valor de la variable aleatoria con respecto al valor esperado. Las ponderaciones corresponden al respectivo valor de la función de probabilidad. La raíz cuadrada positiva de la varianza es conocida como la desviación estándar (σ). La varianza de la variable X de nuestro experimento con los dos dados no cargados es de 79.96, y la desviación estándar 8.9. La desviación estándar es una mejor medida de dispersión, ya que está expresada en las mismas unidades que la variable X.

En las próximas secciones utilizaremos la media y la desviación estándar para caracterizar cierto tipo de distribuciones.

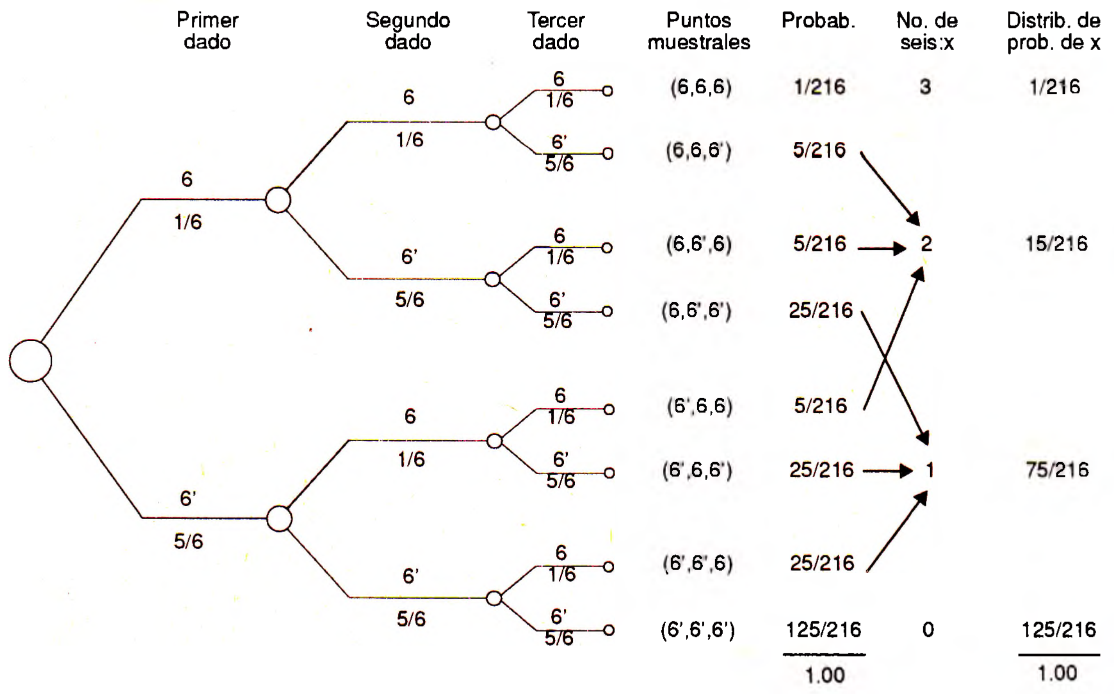
5. LA DISTRIBUCIÓN BINOMIAL

Existen muchas aplicaciones que pueden ser descritas usando ciertas variables aleatorias discretas. En esta y en la próxima sección se presentan dos distribuciones que ayudan a resolver problemas con variables aleatorias discretas: la distribución binomial y la distribución de Poisson.

Definamos el experimento que consiste en lanzar tres dados no cargados y una nueva variable X que represente el número de seis (6) que obtenemos en los tres dados. El árbol de probabilidades que representa este experimento se muestra en el gráfico 28. Nótese que al construir el árbol no se detallan todos los posibles resultados que se obtienen al lanzar un dado. Estos se han resumido en sólo dos resultados: obtener un seis (6) y no obtener un seis (6′).

Este experimento tiene ocho puntos muestrales con sus respectivas probabilidades de ocurrencia. Existen cuatro valores diferentes de la variable X asociados a los puntos muestrales. Usando la distribución de probabilidades de X podemos contestar a varias preguntas con respecto al número de "6" que podemos obtener en el lanzamiento de tres dados no cargados.

Gráfico 28: Árbol de probabilidades del experimento de lanzar tres dados no cargados



En el caso de experimentos más complejos, como podría ser lanzar un dado siete veces, los árboles de probabilidades resultarán "frondosos", complicando su construcción y utilización.

Lanzar un dado o una moneda muchas veces o detectar el número de piezas defectuosas de un cargamento pertenecen a una clase de experimentos que poseen las siguientes características:

1. El experimento total puede ser descrito en términos de una secuencia de n experimentos idénticos llamados *pruebas*.

2. Todas las pruebas tienen espacios *muestrales dicótomos idénticos*; o, en otras palabras, hay dos resultados posibles en cada prueba. Usualmente, a uno de los resultados se le llama *éxito* y al otro *fracaso*. En nuestro experimento de lanzar el dado tenemos una situación binaria: aparecerá "6" (éxito) o no aparecerá "6" (fracaso).

3. Las *probabilidades* de los dos resultados posibles son constantes en todas las pruebas del experimento.

4. Las pruebas son independientes, es decir que el resultado de una prueba no afecta el de ninguna otra.

Los experimentos que satisfacen las condiciones 2, 3 y 4 se dice que son generados por un *proceso de Bernoulli*. Si además se satisface la condición 1 (hay n pruebas), diremos que tenemos un *experimento binomial*. La variable aleatoria discreta asociada con el experimento binomial es el número de *éxitos* en las n pruebas. Si denotamos con X el valor de esta variable aleatoria, entonces X puede tomar los valores de 0, 1, 2,..., n, dependiendo del número de éxitos observados en las n pruebas. La distribución de probabilidades asociadas con esta variable aleatoria es llamada *distribución de probabilidades binomial*. En los casos en los que la distribución binomial es aplicable se puede usar una fórmula matemática para calcular la probabilidad asociada con cualquier valor de la variable aleatoria. Mostraremos cómo se puede derivar esa fórmula usando nuestro experimento de lanzar tres dados.

Sabemos que la probabilidad de obtener un "6" en un dado no cargado es 1/6. ¿Cuál es la probabilidad de obtener exactamente dos "6" en tres dados no cargados? ¿Cuál es la probabilidad de no obtener ningún "6" en tres dados? En primer lugar vemos que nuestro experimento es binomial: a) El experimento puede ser descrito como una secuencia de tres pruebas idénticas, una prueba para cada uno de los tres dados no cargados; b) Dos resultados son posibles en cada prueba (se obtiene un "6" o no); c) Las probabilidades son las mismas para todos los dados (1/6 para obtener un "6", y 5/6 para no obtener un "6"); y, d) El resultado de un dado es independiente del resultado de los otros dados. Luego la variable X, número de "6" obtenidos al lanzar tres dados, satisface los requerimientos de la distribución de probabilidades binomial.

Ahora tratemos de determinar la probabilidad de obtener exactamente dos "6" en los tres dados no cargados. En el gráfico 28 vemos que hay tres resultados posibles en los cuales ocurren dos "6": (6, 6, 6′), (6, 6′, 6) y (6′, 6, 6). Dado que las pruebas son independientes y las probabilidades constantes, podemos desarrollar una expresión para determinar la probabilidad de cada uno de los puntos muestrales donde existen dos éxitos en las tres pruebas. Tomemos, por ejemplo, el punto muestral (6,6,6′), puesto que la probabilidad de éxito es (1/6) y es constante en todas las pruebas, siendo la probabilidad de fracaso de (5/6). La probabilidad de este punto será:

$$p(6,6,6'/e) = (1/6)\,(1/6)\,(5/6) = (1/6)^2\,(5/6)^1 = 5/216 \qquad (11)$$

De igual manera podemos calcular la probabilidad para los otros dos resultados con dos éxitos: (6,6′,6) y (6′,6,6). Como se aprecia en el gráfico 28, estas probabilidades son iguales a la probabilidad del punto (6,6,6′). Entonces la probabilidad de obtener exactamente dos "6" en tres dados será:

$$p(\text{Probabilidad de obtener dos "6" al lanzar tres dados}) = 3 * (1/6)^2\,(5/6)^1 = 15/216$$

Podemos desarrollar una *expresión general* para determinar la probabilidad de X éxitos en un experimento binomial con n pruebas. Si p representa la probabilidad de éxito y (1 – p) la probabilidad de fracaso en una prueba, y dado un experimento binomial con n pruebas, la probabilidad de *cada punto muestral* con exactamente X éxitos está dada por:

$$p^{x}\,(1 - p)^{n-x} \qquad (12)$$

El exponente de p representa el número de éxitos, y el exponente de (1 - p) el número de fracasos.

Dado que habrá más de un resultado del experimento con X éxitos en las n pruebas, debemos multiplicar la expresión (12) por el número de puntos muestrales que den exactamente X éxitos en n pruebas. Este número puede calcularse utilizando la fórmula que permite determinar el total de maneras diferentes en que se pueden combinar n "artículos" tomados en grupos de tamaño X. Esta fórmula es:

$$\binom{n}{x} = {}_{n}C_{x} = \frac{n\,!}{x\,!(n-x)\,!} \qquad (13)$$

Para el experimento binomial, esta fórmula de las combinaciones nos brinda el número de resultados del experimento con X éxitos en n pruebas.

Entonces, la probabilidad de obtener X éxitos en n pruebas en un experimento binomial estará dada por el producto de los términos (12) y (13). Por tanto, la fórmula matemática para la función de distribución binomial es la siguiente:

$$p_b\,(x/n,\,p) = \frac{n!}{x!\,(n-x)\,!}\; p^{x}\,(1-p)^{n-x} \qquad (14)$$

donde p_b (x/n, p) es la probabilidad de que ocurran X éxitos en n pruebas posibles, dada la probabilidad de éxito p.

Para nuestro ejemplo de los tres dados no cargados, n = 3, x = 2 y p = 1/6. La probabilidad de obtener exactamente dos éxitos al lanzar tres dados usando la ecuación (14) es:

$$p_b(x = 2/n = 3, p = 1/6) = \frac{3!}{2!\,1!}(1/6)^2(5/6)^1 = 15/216$$

Este valor es idéntico al calculado utilizando el árbol de probabilidades.

La ecuación (14) muestra que si un experimento posee las cuatro propiedades binomiales, se puede calcular la probabilidad de cualquier número particular de éxitos (X) con sólo conocer el número de pruebas (n) y la probabilidad de éxito en cada prueba (p). También se pueden calcular muy fácilmente las distribuciones de probabilidades acumuladas:

$$p_b(X \leq X_0/n, p) = \sum_{X \leq X_0} \frac{n!}{X!\,(n-X)!} p^x (1-p)^{n-x} \qquad (15)$$

La distribución de probabilidades acumuladas complementaria:

$$p_b(X > X_0/n, p) = 1 - p_b(X \leq X_0/n, p) \qquad (16)$$

Las distribuciones de probabilidades de la variable aleatoria X, definida como el número de "6" obtenidos al lanzar tres dados no cargados, se muestran en el cuadro 10. Para el cálculo de estas distribuciones se usan las ecuaciones (14), (15) y (16).

Note que los resultados presentados en el cuadro 10 para $p_b(X = X_0)$ son los mismos que los obtenidos usando el árbol de probabilidades del gráfico 28. En general, se puede usar el árbol de probabilidades o la función de probabilidades binomial, ecuación (14), para determinar la probabilidad de X éxitos siendo esta última más expedita cuando el número de pruebas se incrementa.

CUADRO 10: DISTRIBUCIONES DE PROBABILIDADES DE X, EL NÚMERO DE "6" OBTENIDOS AL LANZAR TRES DADOS NO CARGADOS

X	$p_b(X = X_0)$	$p_b(X \leq X_o)$	$p_b(X > X_o)$
0	125/216	125/216	91/216
1	75/216	200/216	16/216
2	15/216	215/216	1/216
3	1/216	216/216	0
Total	1		

Caracterización de la distribución binomial

Usando la definición de la media o esperanza matemática de una variable aleatoria, tenemos:

$$\mu_b = E(X) = \sum x\, p(X/e) = \sum x\, p_b(x) \qquad (17)$$

Sustituyendo P_b (X) por la ecuación (14), y simplificando:

$$\mu_b = n\, p \qquad (18)$$

De igual manera, podemos encontrar una fórmula simple para la varianza y desviación estándar de la distribución binomial, utilizando las ecuaciones (10) y (14):

$$\begin{aligned} \sigma_b^2 &= \sum (x - u)^2\, P_b(X) \\ &= n\, p\, (1-p) \qquad (19) \end{aligned}$$

y,

$$\sigma_b = \sqrt{n\, p\, (1 - p)} \qquad (20)$$

Retomando nuestro ejemplo, usaremos las fórmulas (18) y (20) para calcular el valor esperado y la desviación estándar del número de "6" que se obtienen al lanzar tres dados no cargados:

$$\mu_b = n\,p = 3\,(1/6) = 0.5$$

$$\sigma_b = \sqrt{n\,p\,(1-p)} = \sqrt{3\,(1/6)\,(5/6)} = 0.645$$

6. LA DISTRIBUCIÓN DE POISSON

La función de distribución de Poisson fue derivada como un caso límite de la distribución binomial cuando hay un gran número de pruebas (n) y la probabilidad de éxito (p) es muy pequeña en cada una de ellas:

$$p_p\,(X/\mu) = \frac{\mu^x\,e^{-\mu}}{x\,!} \qquad (21)$$

donde μ es el promedio de éxitos en un intervalo
X es el número de éxitos
e es igual a 2.71828, base de los logaritmos naturales

La ecuación (14) tiende a la ecuación (21) cuando n se aproxima a infinito y p se aproxima a cero.

Si bien la distribución de Poisson fue originalmente derivada de la distribución binomial, posteriormente se ha convertido en una distribución por derecho propio. La ecuación (21) provee un modelo para describir el número de "ocurrencias" de un evento (éxitos) en un intervalo específico de tiempo o espacio. Por ejemplo, la variable aleatoria de interés podría ser el número de accidentes en la carretera Lima-Pucusana un domingo de verano en una hora específica; el número de

piezas defectuosas que produce cierta máquina en una hora; el número de clientes que llegan a un supermercado en una hora, etcétera.

Los supuestos implícitos en la distribución de Poisson son: a) La probabilidad de ocurrencia del evento es la misma en dos intervalos de la misma longitud de tiempo o espacio; y, b) La ocurrencia o no ocurrencia del evento en cualquier intervalo es independiente de su ocurrencia en cualquier otro intervalo.

Para usar la distribución binomial se requiere conocer tanto el número de pruebas (n) como la probabilidad de éxito (p). Existen algunos casos en los cuales se conoce lo que sucedió en promedio en el pasado, pero no se sabe el número de pruebas y/o la probabilidad de éxito en una prueba. Supongamos que deseamos averiguar la probabilidad de que ocurra un número X de accidentes en una hora determinada en la carretera Lima-Pucusana, un domingo de verano. Para utilizar la distribución binomial necesitaríamos saber el número de viajes (n) realizados en dicho intervalo y la probabilidad (p) de que suceda un accidente en cada viaje. Sin embargo, la distribución de Poisson sólo requiere conocer el número promedio de accidentes (μ) ocurridos en una hora, para predecir la probabilidad de x accidentes ($x = 0, 1, 2,...$).

Además, la distribución de Poisson es un modelo útil para describir la probabilidad de eventos raros, cuando p tiende a cero.

Caracterización de la distribución de Poisson

Al igual que en la distribución binomial, podemos derivar las expresiones para la media y la desviación estándar de la distribución de Poisson usando la definición de esperanza matemática y varianza de una variable aleatoria:

$$E(X) = \Sigma X\, p_p(X)$$

$$= \Sigma X \frac{\mu^x e^{-\mu}}{X!}$$

$$E(X) = \mu \qquad (22)$$

$$Var(X) = \Sigma (x-\mu)^2\, p_b(X)$$

$$= \Sigma (x-\mu)^2 \frac{\mu^x e^{-\mu}}{X!}$$

$$Var(X) = \mu \qquad (23)$$

y la desviación estándar será:

$$\sigma_p = \sqrt{\mu} \qquad (24)$$

En la distribución de Poisson, la media y la varianza toman el mismo valor (μ).

Cuando en una distribución de Poisson los valores de X son muy grandes, el cálculo de las probabilidades usando la fórmula (21) puede tornarse muy engorroso dado que hay que evaluar (μ^X) y (e^{-u}). Para evitar hacer estos cálculos se usan tablas estadísticas que contienen las probabilidades de la distribución de Poisson para diferentes valores de μ y X (ver apéndice A). En estas tablas los diferentes valores de μ van en la fila superior, y las probabilidades para cada valor de X están en las columnas debajo del valor correspondiente de μ, p(X/μ). El siguiente ejemplo ilustra el uso de estas tablas.

Suponga que estamos interesados en la probabilidad de que llegue un número determinado de personas al consultorio dental de un policlínico situado en un distrito populoso de Lima, durante un período de 30 minutos el sábado por la mañana. Si suponemos que la probabilidad de que llegue un paciente es la

misma en dos intervalos iguales de tiempo, y que la llegada o no de un paciente en cualquier intervalo es independiente de la llegada o no en cualquier otro, entonces la distribución de Poisson es aplicable. Además, la información histórica del policlínico muestra que el número promedio de pacientes que llegan en 30 minutos un sábado en la mañana es 10. Podemos recurrir a las tablas y encontrar la probabilidad para diferentes valores de X, siendo μ igual a 10. Así, si el Administrador del policlínico quiere saber cuál es la probabilidad de que lleguen exactamente siete pacientes en 30 minutos, debemos localizar este valor en la fila de la tabla correspondiente a $x = 7$, y la columna correspondiente a $\mu = 10$. El valor que muestra la tabla para $p_P(x = 7/\mu = 10)$ es 0.0901, que indica una probabilidad de 9.01 % de que lleguen siete pacientes en 30 minutos.

La distribución de Poisson como una aproximación de la binomial

Se puede demostrar que a medida que n se hace grande y pequeño, la distribución binomial se aproxima a la distribución de Poisson con $\mu = np$. Esta es la razón por la cual esta última provee una buena aproximación de la distribución binomial. Una *regla práctica* es que la aproximación de Poisson será buena siempre que $p \leq 0.05$ y $n > 20$. Puesto que los cálculos que se requieren para determinar los valores de la distribución de Poisson son más sencillos que los requeridos para la distribución binomial, esta aproximación se usa con frecuencia.

Para ilustrar el uso de la distribución de Poisson como aproximación de la binomial, consideremos el caso de una máquina que produce piezas con un promedio de 1% de piezas defectuosas. El Supervisor de la planta está interesado en saber la probabilidad de que ninguna pieza sea defectuosa en una muestra de 20. Usando la distribución binomial, encontramos que:

$$p(x = 0/n = 20, p = 0.01) = 0.8179.$$

Para poder usar la distribución de Poisson establecemos que:

$$\mu = np = (20)(.01) = 0.2$$
$$x = 0$$

Las tablas de Poisson muestran que:

$$p(x = o/\mu = 0.2) = 0.8187$$

Vemos que la diferencia entre las probabilidades calculadas con la distribución binomial y Poisson es mínima. Para valores mayores de n y menores de p la aproximación será más exacta.

7. VARIABLES ALEATORIAS CONTINUAS: LA DISTRIBUCIÓN NORMAL

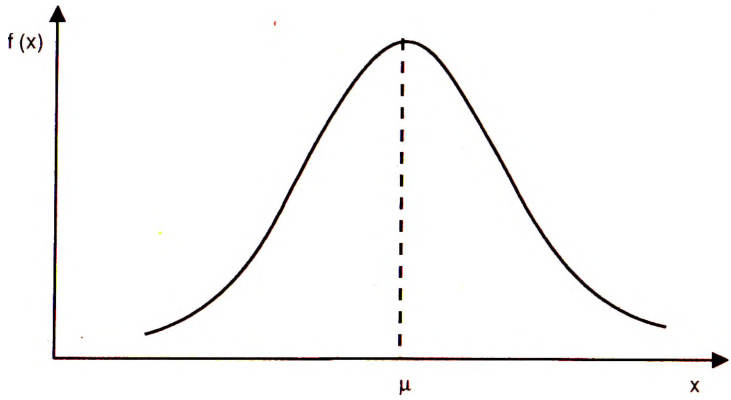
Consideremos las siguientes variables aleatorias: el peso del arroz contenido en una bolsa de "dos kilos", la temperatura en cierto momento del día, el tiempo de vuelo de un avión entre Lima y Arequipa, la vida útil de un foco eléctrico de 50 watts. Los valores que pueden tomar estas variables no están circunscritos a un número finito de posibilidades, sino que pueden variar en forma continua en un intervalo dado. Estas variables se denominan variables aleatorias continuas.

Para entender mejor la naturaleza de las variables aleatorias continuas, consideremos el ejemplo de las bolsas de arroz de "dos kilos". Por muy calibrada que se encuentre la máquina embolsadora de arroz, es difícil que cada bolsa pese exactamente dos kilos. Existirán un número infinito de valores alrededor de los dos kilos que representan el peso verdadero de dichas bolsas. Es imposible listar cada uno de los pesos exactos de arroz en cada bolsa que constituyen los valores de esta variable aleatoria, e identificar la probabilidad asociada con ellos. De hecho, para las variables aleatorias continuas necesitamos introducir un

nuevo método para calcular las probabilidades asociadas con los valores posibles de la variable aleatoria.

En esta sección sólo se presentará la distribución de probabilidades más importante para describir una variable aleatoria continua conocida como la *distribución de probabilidades normal*. La distribución normal tiene muchos usos en el análisis estadístico, ya que puede describir una serie de situaciones prácticas como las mencionadas anteriormente. Además, la distribución normal es la base de la estadística inferencial y del análisis de regresión, como se verá en los capítulos siguientes. La función que describe la distribución de probabilidades normal es conocida como *función de densidad*, y tiene una forma muy similar a la de una campana, como se muestra en el gráfico 29.

Gráfico 29: Función de densidad de una distribución normal con media μ



La expresión matemática que describe la curva en forma de campana de la función de densidad normal es la siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (25)$$

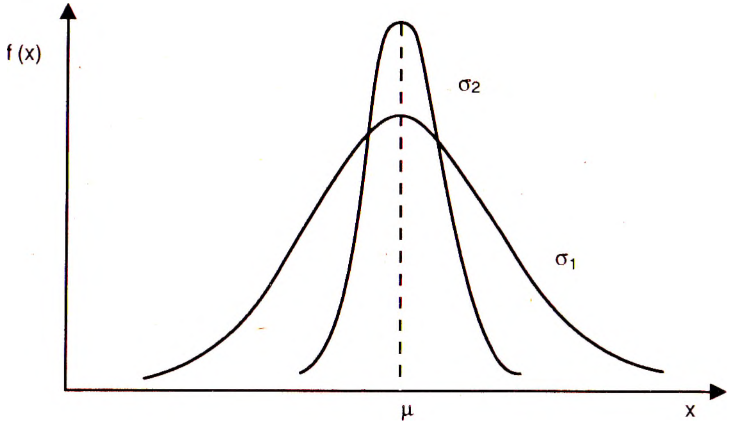
donde	μ	es la media o valor esperado de la variable aleatoria X
	σ^2	es la varianza de la variable aleatoria X
	σ	es la desviación estándar
	π	es igual a 3.1416
	e	es igual a 2.7183

En la representación gráfica de la función de densidad, f(x) muestra la altura o valor de la función para cada valor particular de X. Una vez que se ha calculado la media (μ) y la varianza (σ^2), la ecuación (25) puede usarse para graficar la distribución normal correspondiente. Si tenemos dos distribuciones normales con igual media pero con diferentes desviaciones estándar, la curva que representa a la distribución con mayor desviación estándar será más "achatada"; es decir, los valores de la variable no estarán tan concentrados alrededor de la media (ver gráfico 30).

A diferencia de la distribución de probabilidades de una variable aleatoria discreta, la función de densidad representa la altura de la función en el valor particular de X y no su probabilidad de ocurrencia. Recordemos que para cada valor de una variable aleatoria discreta, la distribución de probabilidades nos señala la probabilidad de que X tenga exactamente ese valor. Sin embargo, dado que la variable aleatoria continua tiene un número infinito de posibles valores, ya no podemos tratar de identificar la probabilidad para cada valor específico de X. Más bien debemos considerar la probabilidad sólo en términos de la posibilidad de que la variable aleatoria continua tome un valor en un intervalo específico. Esto está dado por el área bajo la curva normal f(x). Una vez que f(x) ha sido identificada para una variable aleatoria

continua, entonces la probabilidad de que X se encuentre entre un valor "a" y otro mayor "b" puede ser determinada calculando el área bajo la curva de f(x) en el intervalo entre "a" y "b".

Gráfico 30: Funciones de densidad de dos distribuciones normales con desviaciones estándar $\sigma_1 > \sigma_2$ y con la misma media μ



Debemos notar que el área total bajo la curva f(x) es igual a 1, para cualquier variable aleatoria continua. Esta propiedad es análoga a la condición de que la suma de las probabilidades de una distribución de probabilidades discreta debe ser igual a uno. Para una distribución de probabilidad continua también se requiere que $f(x) \geq 0$ para todos los valores de X, para cumplir con el primer axioma de la teoría de probabilidades.

Al igual que con las variables aleatorias discretas, las continuas pueden también caracterizarse a través de su valor esperado E(x) y su varianza Var(x). El cálculo de estos parámetros requiere del cálculo integral:

$$E(x) = \int x\, f(x)\, dx$$

$$Var(x) = \int [x - E(x)]^2 f(x)\, dx$$

En el caso de la distribución normal, la función de densidad depende de la media (μ) y de la desviación estándar (σ). Para encontrar la solución de problemas que involucran variables con una distribución normal, no tenemos que usar necesariamente la función de densidad especificada por la ecuación (25). Podremos recurrir a tablas estadísticas que proporcionan valores de probabilidades o áreas bajo la curva f(x), para una distribución normal estandarizada.

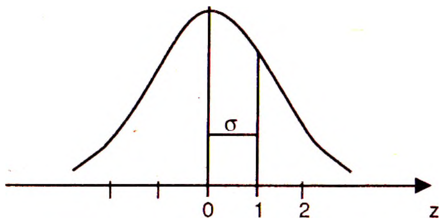
A. La distribución normal estandarizada

Una variable aleatoria que tiene una distribución normal con una media igual a cero y una desviación estándar igual a uno, se dice que tiene una distribución normal estandarizada. Se usa la letra "z" para denotar a esta variable normal específica. Su gráfico tiene la misma forma que otras distribuciones normales, con las propiedades especiales de $\mu = 0$ y $\sigma = 1$ (ver gráfico 31). Tomando como base estos parámetros se han calculado y tabulado las probabilidades para la distribución normal estandarizada. Esta tabla se presenta en el apéndice B, y nos permite determinar la probabilidad de que el valor de la variable normal estandarizada se encuentre en un intervalo específico, entre la media (z = 0) y cualquier valor positivo de z; es decir, nos permite establecer el área bajo la curva normal en dicho intervalo. En términos de la distribución de probabilidades acumulada, diremos que esta tabla nos da los valores de la probabilidad:

$$p(0 \leq z \leq z_0)$$

para valores no negativos de z.

Gráfico 31: Función de densidad de la distribución normal estandarizada, z



La tabla de la distribución normal estandarizada presenta los valores de z con una aproximación a décimos en la primera columna de la izquierda. El segundo decimal de z aparece en la fila superior. Por ejemplo, para z = 1.25 encontraremos 1.2 en la columna de la izquierda y 0.05 en la fila superior. Luego, el valor del área debajo de la curva entre la media (z = 0.00) y z = 1.25 será 0.3944. En términos de la distribución de probabilidades acumulada, tenemos:

$$p(0 \leq z \leq 1.25) = 0.3944$$

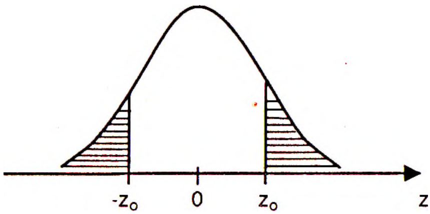
Tomemos otro ejemplo: El área o probabilidad de que z esté en el intervalo de z = 0.00 a z = 1.50 está dada por:

$$p(0 \leq z \leq 1.50) = 0.4332$$

Los valores para la distribución de probabilidades acumuladas para valores negativos de z pueden determinarse dado que la función de densidad de la distribución normal es simétrica alrededor de su media. Como se ilustra en el gráfico 32, el área bajo la curva a la izquierda de - Z_0 es la misma que el área bajo la curva a la derecha de Z_0, es decir:

$$p(z \leq -z_0) = p(z \geq z_0)$$

Gráfico 32: Función de densidad de la distribución normal estandarizada, z p $(z \leqslant -z_o) = p\ (z \geqslant z_o)$



Por otro lado, sabemos que el área bajo la curva a la derecha de la media es 0.5; entonces:

$$p(z \leq -z_0) = p(z \geq z_0) = 0.5 - p(0 \leq z \leq z_0)$$

y el valor de $p(0 \leq z \leq z_0)$ está dado en la tabla del apéndice B. Por ejemplo, si deseamos saber la probabilidad de que z sea menor o igual a -1.5:

$$p(z \leq -1.5) = 0.5 - p(0 \leq z \leq 1.5) = 0.5 - 0.4332 = 0.0668$$

De igual manera podemos calcular las probabilidades para otros intervalos de z. Por ejemplo, si deseamos la probabilidad de que z esté entre -1.5 y 1.25, es decir $p(-1.5 \leq z \leq 1.25)$, podemos calcularla de la siguiente manera:

$$\begin{aligned} p(-1.5 \leq z \leq 1.25) &= p(-1.5 \leq z \leq 0) + p(0 \leq z \leq 1.25) \\ &= p(0 \leq z \leq 1.5) + p(0 \leq z \leq 1.25) \\ &= 0.4332 + 0.3944 \\ &= 0.0876 \end{aligned}$$

B. Cálculo de probabilidades para cualquier distribución normal

Cualquier distribución normal puede ser convertida a la distribución normal estandarizada para facilitar el cálculo de sus probabilidades. La transformación de una variable aleatoria normal, X, con media μ y desviación estándar σ, a la variable normal estandarizada, z, se efectúa de la siguiente manera:

$$z = \frac{x - \mu}{\sigma} \qquad (26)$$

Entonces z mide la desviación de X respecto de la media, μ, en unidades de desviación estándar.

Con esta transformación, la función de densidad de la ecuación (25) se convierte en:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-1/2\, z}, \quad \text{con } \mu_z = 0, \sigma_z = 1$$

Para ilustrar la conversión de una variable aleatoria normal, X, a una variable normal estandarizada, z, consideremos el siguiente ejemplo: Un taladro debe hacer orificios de 1 pulgada de diámetro en placas de hierro. La máquina se ajusta de modo que haga una perforación de 1.02 pulgadas de diámetro con el fin de asegurar que el orificio no tenga menos de 1 pulgada de diámetro. Si la desviación estándar del diámetro de los orificios hechos con esta máquina es 0.01, y se supone que los diámetros de las perforaciones se distribuyen normalmente, ¿cuál es el porcentaje de orificios que tendrán menos de 1 pulgada de diámetro?

Tenemos una variable aleatoria con una distribución normal con $\mu = 1.02$ y $\sigma = 0.01$. Esta distribución se muestra en el gráfico 33. Note que aparte de los valores de la variable aleatoria X,

hemos incluido un segundo eje horizontal (z) para mostrar que para cada valor de X hay un valor correspondiente de z. Así, cuando x = 1.02 el valor correspondiente de z es:

$$z = \frac{x - \mu}{\sigma} = \frac{1.02 - 1.02}{0.01} = 0$$

De manera similar, para X = 1.00 tenemos:

$$z = \frac{x - \mu}{\sigma} = \frac{1.00 - 1.02}{0.01} = -2$$

¿Cuál es, entonces, el porcentaje de orificios que tendrán menos de 1 pulgada de diámetro? Es decir:

$$p\,(x < 1) = ?$$

No tenemos tablas que nos den esta probabilidad directamente. Sin embargo, notemos que en el gráfico 33 el área bajo la curva para X < 1 es la misma que el área bajo la curva para z < -2.

$$p\,(x < 1) = p\,(z < -2)$$

Usando la tabla de la distribución normal estandarizada, encontramos que el área o probabilidad de que z sea menor que -2 es:

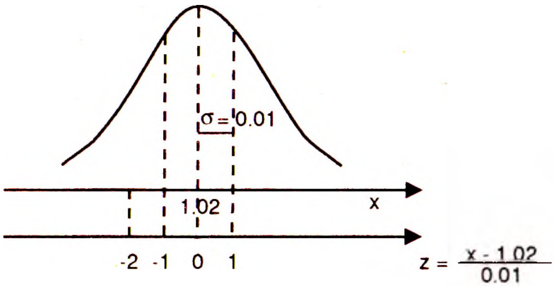
$$\begin{aligned} p(z < -2) = p\,(z > 2) &= 0.5 - p(0 \leq z < 2) \\ &= 0.5 - 0.4772 = 0.0228 \end{aligned}$$

Concluimos que el 2.28% de los orificios que perfora la máquina tendrá un diámetro menor que 1 pulgada.

El procedimiento descrito se aplica para cualquier problema que involucra variables aleatorias con una distribución normal.

Es decir, para cada valor de X hay un valor correspondiente a z dado por la ecuación (26). Para encontrar la probabilidad de que X esté en un intervalo específico, simplemente transformamos el intervalo X a su correspondiente intervalo z, y lo buscamos en las tablas de probabilidades de la distribución normal estandarizada.

Gráfico 33: Función de densidad de la distribución normal con media 1.02, desviación estándar 0.01, y su conversión a la variable normal estandarizada, z



● *Ejercicios*

1. Un nuevo tratamiento para la inflamación muscular conocida como el "codo del tenista" ha sido desarrollado y está siendo probado con tres tenistas del Club Lawn Tenis que sufren de este mal. Los resultados del experimento son tres: el paciente no mejora, mejora moderadamente o mejora considerablemente.
a. Represente los resultados del experimento usando un árbol de eventos.
b. Defina una variable aleatoria que represente el número de pacientes que no mejoran. Establezca el valor que la variable aleatoria tendrá para cada uno de los resultados del experimento.

2. Basado en información generada por tratamientos similares al presentado en el problema anterior, el doctor Graña, tenista empedernido, considera que existe una probabilidad del 0.30 de que un paciente no mejore, y que sólo existe una probabilidad del 0.20 de que tenga una mejora significativa, cuando se le aplica el nuevo tratamiento.

a. Establezca las distribuciones de probabilidades para la variable definida en la parte (b) del problema anterior.

b. Dibuje los diagramas para las distribuciones establecidas.

c. ¿Cuál es la probabilidad de que ninguno de los tres pacientes tratados mejoren?

d. ¿Cuál es la probabilidad de que por lo menos uno tenga alguna mejoría?

3. Aníbal Pérez, un corredor de bienes raíces, pone un anuncio en el periódico *El Comercio* que describe un departamento que está tratando de vender. De su experiencia pasada, él tiene los siguientes datos acerca de las llamadas telefónicas recibidas en respuesta a anuncios similares en el periódico:

Número de llamadas	Probabilidad de este número de llamadas	Probabilidad de una venta, dado este número de llamadas
6 ó más	0.04	0.95
5	0.12	0.80
4	0.16	0.70
3	0.18	0.50
2	0.27	0.20
1	0.19	0.05
Ninguna	0.04	0.00

a. ¿Cuál es el número esperado de llamadas?

b. ¿Cuál es la varianza en el número de llamadas?

c. ¿Cuál es la probabilidad de que Aníbal venda esta propiedad mediante este anuncio?

d. Si Aníbal vende la propiedad, ¿cuál es la probabilidad de que tuviera dos llamadas?

4. Debido a que no se conoce la estructura geológica de una cierta montaña, el costo de construir un túnel para una carretera es una variable aleatoria descrita de la siguiente manera:

Costo, X (US$)	Probabilidad de que el costo sea X
80 mill.	0.2
100 mill.	0.4
120 mill.	0.3
140 mill.	0.1

a. ¿Cuál es el valor esperado del costo del túnel?
b.¿Qué oferta debe presentarse si el contratista desea estar 90% seguro de que el costo no excederá el ingreso?

5. Existen dos analgésicos en el mercado: "Asperum" y "Alivio". Ambos fabricantes aseguran en sus campañas de publicidad que eliminan el dolor de cabeza más rápidamente que el producto de la competencia. Una agencia independiente de estudios de mercado ha recolectado la siguiente información:

Minutos transcurridos después de tomar el analgésico	Porcentaje de dolores de cabeza eliminados usando: "Asperum"	«Alivio»
10	0.60	0
15	0	0.10
20	0	0.75
25	0	0.15
30	0	0
35	0.40	0

a. ¿Cuáles son las medias y varianzas del tiempo transcurrido para eliminar el dolor de cabeza usando "Asperum" y "Alivio"?
b. ¿Qué analgésico tiene una probabilidad mayor de aliviar un dolor de cabeza en el lapso de 10 minutos?
c. ¿Qué analgésico tiene una probabilidad mayor de aliviar un dolor de cabeza en el lapso de 20 minutos?

6. La demanda mensual de uno de los productos de DetuPerú varía grandemente de un mes a otro. Con base en la información de los

últimos 24 meses, se logró estimar la distribución de probabilidades para la demanda mensual del producto bajo estudio.

Demanda (unidades)	20,000	30,000	40,000	50,000
Probabilidad	0.25	0.35	0.30	0.10

a. Si la compañía establece su programa de producción tomando como base el valor esperado de la demanda mensual, ¿cuál debería ser el programa mensual de producción para este producto?
b. Suponga que cada unidad vendida produce un ingreso de $8 y que su costo es de $6. ¿Cuanto ganará o perderá DetuPerú en un mes si su programa de producción se basa en su respuesta en (a) y la demanda fuera de 20,000 unidades?

7. Un proceso industrial produce ciertas piezas que son clasificadas bien como aceptables, bien como defectuosas. Si la probabilidad de que el proceso produzca una pieza defectuosa es 0.10:
a. ¿Cuántas piezas defectuosas esperaría encontrar en un lote de 500 piezas?
b. ¿Cuál es la varianza del número de partes defectuosas en el lote?

8. Los registros de mantenimiento revelan que solamente una de cada cien computadoras personales (PC) de cierta marca requieren una reparación durante el primer año de uso. El Administrador de la Universidad del Pacífico ordenó la compra de diez PC de esta marca.
a. Encuentre la probabilidad de que ninguna de las diez PC requieran una reparación durante el primer año de uso.
b. Encuentre la probabilidad de que dos PC requieran una reparación durante el primer año de uso.
c. Encuentre el número esperado de PC que requieran una reparación durante el primer año de uso.
d. Encuentre la varianza.

9. Un vendedor a domicilio visita a ocho clientes potenciales al día. De experiencias pasadas se sabe que la probabilidad de que un cliente potencial efectúe una compra es de 0.10.
a. ¿Cuál es la probabilidad de que un vendedor realice dos ventas en un día?
b. ¿Cuál es la probabilidad de que realice al menos tres ventas en un día?
c. ¿Qué porcentaje de días el vendedor no realiza ninguna venta?
d.¿Cuál es el número esperado de ventas al día? En un período de cinco días, ¿cuántas ventas se espera realizar?

10. En una universidad de la capital se ha encontrado que el 20% de los estudiantes se retiran del curso de Análisis Estadístico. En este semestre se han matriculado 20 estudiantes.
a. ¿Cuál es la probabilidad de que dos o menos estudiantes se retiren?
b. ¿Cuál es la probabilidad de que se retiren exactamente cuatro?
c. ¿Cuál es la probabilidad de que se retiren más de tres?
d. ¿Cuál es el número esperado de retiros?

11. Una compañía estima que la probabilidad de tener problemas disciplinarios con sus empleados en un día particular es de 0.08.
a. ¿Cuál es la probabilidad de que la compañía tenga una semana sin problemas disciplinarios?
b. ¿Cuál es la probabilidad de que tenga exactamente dos días con problemas disciplinarios en dos semanas de trabajo?
c. ¿Cuál es la probabilidad de que al menos tenga dos días sin problemas disciplinarios en cuatro semanas de trabajo?

12. La pizzería "La Romana" es una pizzería muy exclusiva que sólo atiende a diez parejas por noche. "La Romana" es muy famosa por su "Pizza Romana para dos", cuya masa debe ser preparada el día anterior. La probabilidad de que una pareja cualquiera ordene una "Pizza Romana" es 0.40, y cada pareja ordena independientemente de otras parejas.
a. ¿Cuál es el número esperado de "Pizzas Romanas para dos" que se sirven por noche? ¿Cuál es la varianza?
b. ¿Cuántas "Pizzas Romanas" deben prepararse si se quiere que la probabilidad de no tener "Pizzas Romanas" suficientes para todos los clientes que la deseen no sea mayor que 0.10?

13. De 800 familias con cuatro hijos cada una, ¿qué porcentaje se espera que tengan:
a. 2 niños y 2 niñas?
b. al menos un niño?
c. ninguna niña?
d. a lo más 2 niñas?

Suponga que la probabilidad de nacimiento de un niño es la misma que la de una niña.

14. Si el 30% de los estudiantes de la Escuela de Postgrado tiene visión defectuosa, ¿cuál es la probabilidad de que por lo menos la mitad de los miembros de una clase de veinte estudiantes posea visión defectuosa? Utilizar la aproximación de la distribución de Poisson.

15. Durante las horas de mayor tráfico en Lima Metropolitana, ocurren en promedio dos accidentes por hora. En las mañanas el

período de mayor tráfico dura 1 hora y 30 minutos, mientras que en la tarde tiene una duración de 2 horas.

a. En un día en particular, ¿cuál es la probabilidad de que no haya ningún accidente en el período "pico" de la mañana?

b. ¿Cuál es la probabilidad de que sucedan dos accidentes durante el período "pico" de la tarde?

c. ¿Cuál es la probabilidad de que ocurran tres o más accidentes en el período "pico" de la mañana?

16. En el período de matrícula de la Escuela de Postgrado de la Universidad del Pacífico, los estudiantes consultan a los profesores acerca de la elección de cursos a tomar. Un profesor se dio cuenta de que en este período un promedio de diez estudiantes por hora se acercaban a consultarle; sin embargo, el número exacto de estudiantes que le consultaban era aleatorio. Use la distribución de Poisson para contestar a las siguientes preguntas:

a. ¿Cuál es la probabilidad de que exactamente diez estudiantes consulten a este profesor en una hora del período de matrícula?

b. ¿Cuál es la probabilidad de que tres estudiantes consulten a este profesor en un lapso de media hora?

17. Un nuevo proceso de producción automático ha experimentado un promedio de 2.5 interrupciones por día. Debido a los costos asociados a una interrupción, el Administrador está preocupado con respecto a la posibilidad de tener dos o más interrupciones durante un día específico. Suponga que el número de interrupciones por día sigue una distribución de Poisson. ¿Cuál es la posibilidad de tener dos o más interrupciones?

18. Las llegadas de los clientes a la tienda Wong de San Isidro siguen una distribución de Poisson. Suponga que la llegada promedio de clientes es de tres por minuto.

a. ¿Cuál es la probabilidad de que lleguen exactamente tres clientes en un minuto?

b. ¿Cuál es la probabilidad de que lleguen al menos tres clientes en un minuto?

19. Una compañía produce focos de luz cuyo tiempo de vida sigue una distribución normal con una media de 1,200 horas y una desviación estándar de 250 horas. Un foco se selecciona aleatoriamente de la producción de la compañía:

a. Encuentre la probabilidad de que dure menos de 1,500 horas.

b. Encuentre la probabilidad de que el foco seleccionado dure al menos 1,000 horas.

c. Encuentre la probabilidad de que dure entre 1,000 y 1,400 horas.

d. Sin hacer los cálculos necesarios, determine en cuál de los siguientes rangos es más probable que se encuentre su duración:

(i)	1,000 - 1,200 horas
(ii)	1,100 - 1,300 horas
(iii)	1,200 - 1,400 horas
(iv)	1,300 - 1,500 horas

¿Por qué?

20. La hora en la cual el cartero acostumbra repartir el correo en la Universidad del Pacífico sigue una distribución normal. La hora promedio de reparto es a las 10:00 a.m., con una desviación estándar de 15 minutos.
a. ¿Cuál es la probabilidad de que la correspondencia llegue después de las 10:30 a.m.?
b. ¿Cuál es la probabilidad de que la correspondencia llegue antes de las 9:36 a.m.?
c. ¿Cuál es la probabilidad de que la correspondencia llegue entre las 9:48 y las 10:09 a.m.?

21. Considere la variable X como el tiempo entre llegadas de pacientes a un hospital. Esta variable tiene una media de 20 minutos y una desviación estándar de 8 minutos. Suponga que X se distribuye normalmente y encuentre lo siguiente:
a. La probabilidad de que X sea mayor que 2/3 de hora.
b. Las posibilidades son de 2 contra 1 de que X caerá dentro de un intervalo simétrico alrededor de la media. Encuentre el intervalo.

Considere ahora la media de una muestra aleatoria de 81 llegadas (X). Hallar:
c. $E(X)$ y σ_X.
d. $p(X \leq 22.00)$.
e. Un número K tal que $p(20.00 - K \leq X \leq 20.00 + K) = 0.97$.

22. El banco "Crédito Popular" está revisando sus cobros por servicios y su política de pago de intereses en las cuentas corrientes. El banco ha encontrado que el balance promedio diario en las cuentas personales es de $ 55, con una desviación estándar de 15. Además, se encontró que los balances promedio diarios se distribuyen normalmente.
a. ¿Qué porcentaje de las cuentas corrientes de los clientes tienen balances promedio diarios mayores de $ 80?
b. ¿Qué porcentaje de los clientes del banco tienen balances promedio diarios menores de $ 20?

c. El banco está considerando pagar intereses a clientes que tengan balances promedio diarios mayores que cierta cantidad. Si el banco no quiere pagar intereses a más del 15% de sus clientes, ¿cuál es el monto mínimo del balance promedio diario al cual está dispuesto a pagar intereses?

23. Servicios de Consultoría de Inversiones (SCI) nos aconseja comprar acciones de la Cadena de Tiendas Montebello, que se están vendiendo a $ 18 por acción. SCI estima que en un año el precio de una acción de Montebello estará a un precio x, donde x tiene una distribución normal con una media de 20 y una varianza de 4. Usando esta información, encuentre la siguientes probabilidades sobre el valor de las acciones dentro de un año:

a. La probabilidad de que las acciones se vendan a $ 20.
b. La probabilidad de que las acciones se vendan a más de $ 20.
c. La probabilidad de que las acciones se vendan a menos de $ 22.
d. El valor esperado de ganancia por acción.
e. La probabilidad de que las acciones se vendan a menos de $ 18.

24. Suponga que las calificaciones de los exámenes de admisión a la Universidad de Ingeniería se distribuyen normalmente, con una media de 450 (sobre un máximo de 1,000), y una desviación estándar de 100.

a. ¿Qué porcentaje de los postulantes que rindió el examen obtuvo una calificación entre 400 y 500?
b. Suponga que Juan Pérez obtuvo una nota de 630. ¿Qué porcentaje de postulantes obtuvo una mejor nota que Juan? ¿Qué porcentaje obtuvo una nota peor?
c. Si la facultad de Ingeniería Electrónica no acepta postulantes con una calificación menor que 480, ¿qué porcentaje de los postulantes que rindieron el examen serían aceptados por esta facultad?

25. Un taller de mecánica automotriz ha encontrado que el tiempo promedio que se requiere para reparar un automóvil es una variable aleatoria que se distribuye normalmente con una media de 50 minutos y una desviación estándar de 15 minutos.

a. De un total de 360 automóviles, ¿cuántas reparaciones se espera que demorarán menos de una hora?
b. De un total de 400 automóviles, ¿cuántas reparaciones se espera que demoren entre 40 y 65 minutos?
c. ¿El 30% de los automóviles tomarán menos de cuántos minutos para ser reparados?

26. La discoteca "Pura Chicha" tiene dos tiendas. Las ventas de cada tienda siguen una distribución normal. La tienda 1 tiene μ= $ 2,000 por

σ

día y μ = \$ 200 por día, y la tienda 2 tiene σ= \$ 1,900 por día y μ = \$ 400 por día.

a. ¿Qué tienda tiene el promedio de ventas más alto?

b. ¿Cuál es la probabilidad de que las ventas diarias sean mayores que \$ 2,200 para la tienda 1? ¿Para la tienda 2?

c.¿Hay alguna contradicción entre las respuestas a las preguntas (a) y (b)? Explique.

27. Chocolates "Primavera" vende una caja de chocolates de dos kilos. Debido a las imperfecciones en el equipo que fabrica los chocolates, el peso real de la caja de chocolates tiene una distribución normal, con un promedio de dos kilos y una desviación estándar de 200 gramos.

a. ¿Cuál es la probabilidad de que una caja de chocolates pese más de 2 kilos 300 gramos?

b. Las regulaciones municipales requieren que al menos el 80% de todos los productos vendidos tenga el peso establecido. ¿Está "Chocolates Primavera" violando estas regulaciones?

c. ¿Cuál es la probabilidad de que una caja de chocolates pese menos de 1 kilo 800 gramos?

IV. Muestreo y estimación

1. Objetivos del muestreo. 2. Diseños muestrales. 3. Distribuciones muestrales de los estadígrafos. 4. Estimación de parámetros. 5. Intervalos de confianza usando la distribución. 6. Tamaño muestral.

En el capítulo I se discutieron las técnicas estadísticas que permiten la presentación y el resumen de los datos referidos a una población. Vimos cómo estas técnicas facilitan la descripción de las características de la población con algunas medidas compactas. En los capítulos II y III aprendimos la teoría de probabilidades y algunas variables aleatorias importantes. Ahora podemos combinar nuestro conocimiento sobre el manejo de información con la teoría de probabilidades para poder derivar inferencias sobre la población entera en la cual tenemos interés, tomando como base información parcial.

En muchos casos es imposible o innecesario tener información de la población completa. Por ejemplo, si queremos determinar la vida útil media de las llantas de una marca determinada, no sería práctico analizar todas y cada una de las llantas producidas. Dicha medida podría estimarse usando tal vez unas cien llantas. Otro ejemplo muy familiar es el uso de las encuestas para predecir el resultado de elecciones democráticas. Para lograr esto se toma un subconjunto del electorado total y con

base en sus opiniones y preferencias se infiere el comportamiento de toda la población.

Este capítulo pretende explicar en términos generales los principios básicos de la teoría de muestreo. En la primera sección se describen los objetivos del muestreo, y en la segunda se discuten los procedimientos probabilísticos y no probabilísticos del diseño muestral. En la sección sobre distribuciones muestrales se formulan los resultados del teorema del límite central, que permiten establecer las distribuciones muestrales de los estadígrafos: media y proporción. En la sección 4 se presenta la estimación de los parámetros de la población sobre la base de los estadígrafos calculados de los datos muestrales. Luego se presenta el concepto fundamental de intervalo de confianza, y finalmente se discuten las consideraciones que deben tenerse en cuenta para determinar el tamaño muestral óptimo.

1. OBJETIVOS DEL MUESTREO

En muchos problemas, los datos de sólo una parte de la población pueden dar la información necesaria para tomar una decisión o probar una hipótesis referente a toda la población. La parte se llama *muestra*, y el todo se llama *población* o *universo*. El objetivo del *muestreo* es seleccionar una muestra que sea representativa de la población. La determinación del *tamaño correcto* de la muestra, del *método* adecuado de selección de tal manera que la muestra represente a la población y la *estimación* de las características de la población tomando como base la muestra son los temas de este capítulo.

Una característica especial de la población se denomina un *parámetro*; su contraparte en la muestra se llama un *estadígrafo* o *estadístico*. Luego, el objetivo principal del muestreo es inferir acerca del parámetro, que es desconocido, basándose en el estadígrafo, que puede ser medido de la información muestral. Si denotamos una muestra variable de interés por la letra X, entonces las observaciones individuales serán X_1, X_2, X_3, ..., X_N. El

tamaño de la población se representa por N, mientras que el tamaño muestral por n. En términos de notación, vamos a usar letras griegas para denotar parámetros y letras latinas para los estadígrafos:

Características	Parámetro	Estadígrafo o estadístico
Media	μ	$\overline{x}$
Proporción	π	p
Desviación estándar	σ	s

En resumen, vamos a tomar una muestra y a usar las leyes de probabilidades para lograr un estimado de los parámetros de la población total. Pero la pregunta clave es: ¿no sería mejor estudiar a la población total? Podemos señalar tres ventajas de usar una muestra para informarse respecto a una población, aun cuando sea posible tomar el censo completo de la población. En primer lugar, el *tiempo* y el *costo* son las razones más importantes para el muestreo; si la muestra está bien seleccionada puede proporcionar información a tiempo, con una exactitud notable y a bajo costo si se compara con el costo de tomar un censo completo. En segundo lugar, en algunos procesos de muestreo, como en las pruebas industriales, los objetos de la muestra se destruyen durante el proceso de prueba; entonces, no podría hacerse un censo completo. Por último, otra razón para preferir el muestreo es la *mayor exactitud* que se puede lograr en la medición de los datos. Un censo completo puede contener numerosos errores de medición debido a que requiere de una gran cantidad de entrevistadores que podrían no estar adecuadamente capacitados para recolectar la información.

2. DISEÑOS MUESTRALES

Dado que la muestra es una parte de la población que servirá para estimar sus características, el diseño del procedimiento de

selección de los elementos de la muestra es de vital importancia. Buscamos que la muestra represente realmente a la población que deseamos estudiar. Existen diversos diseños y técnicas de muestreo para lograr una buena representatividad, y su selección dependerá del tipo de problema que se esté analizando.

Los diferentes diseños muestrales pueden agruparse en dos grandes categorías: probabilísticos y no probabilísticos. Diremos que una *muestra es probabilística o aleatoria* si cada elemento de la población tiene alguna probabilidad de ser elegido en la muestra, probabilidad que debe ser conocida por el investigador. Por otro lado, una *muestra no probabilística* es aquella en la cual algunos elementos de la población no tienen posibilidades de ser seleccionados, o si la probabilidad de elegir un elemento cualquiera no es conocida o no puede determinarse de antemano.

A. Muestras probabilísticas o aleatorias

Dado que en el diseño de las muestras probabilísticas se conoce la probabilidad de seleccionar cada elemento de la población, entonces el investigador podrá utilizar los diferentes resultados de la teoría de probabilidades para evaluar la confiabilidad de las conclusiones que se obtengan a partir de estas muestras. Por esta razón, las muestras probabilísticas pueden ser objeto de análisis y tratamiento estadístico. En este capítulo estudiaremos en detalle el tipo más común de muestras probabilísticas: la muestra probabilística simple. Además, en el resto de esta sección mencionaremos los otros tipos de muestras.

a. Muestra aleatoria simple

En la selección de una muestra aleatoria simple, cada elemento de la población tiene la misma probabilidad de ser incluido en dicha muestra. El diseño de una muestra aleatoria simple implica tres pasos fundamentales. En primer lugar, es necesario defi-

nir claramente la población que será estudiada, y para ello debemos disponer de una lista de todos los elementos de la población. Esta lista se denomina *marco poblacional o referencial*. Por ejemplo, si deseamos estudiar las características socioeconómicas de los estudiantes de la Universidad del Pacífico, debemos definir si el marco poblacional incluirá a los estudiantes de pregrado, a los de postgrado, o a ambos. El marco poblacional debe contener cada unidad elemental de la población y excluir los duplicados. En segundo lugar, debemos definir la *unidad elemental de muestreo*; en el ejemplo anterior, la unidad elemental es el estudiante individual. Y finalmente, debemos establecer el *método de selección* de los elementos que serán incluidos en la muestra. En nuestro ejemplo de los estudiantes de la Universidad del Pacífico, podríamos obtener una muestra tomando como base a todos los estudiantes que se encuentren en la biblioteca el martes a las 8 de la noche. Este procedimiento no selecciona una muestra aleatoria simple, pues no todos los estudiantes tendrán la misma probabilidad de ser elegidos.

Si limitamos nuestro interés a los estudiantes de postgrado de la Universidad del Pacífico, podemos usar la lista de los registros de la Escuela de Postgrado para definir el marco poblacional. Este consistirá de la información relacionada a los 271 estudiantes matriculados. Una muestra aleatoria simple puede obtenerse escribiendo el nombre de cada estudiante en un pedazo de papel, introduciéndolos en una caja y extrayendo unos cuantos al azar. Este proceso tan elemental se torna engorroso cuando el tamaño poblacional es grande, resultando más conveniente el uso de tablas de números aleatorios que han sido construidas especialmente para tales propósitos. Estas tablas presentan combinaciones de los diez dígitos del sistema decimal, los cuales tuvieron la misma probabilidad de ser seleccionados. La tabla del apéndice D presenta números aleatorios de cinco dígitos. A continuación se muestran las primeras cinco filas de dicha tabla.

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	93263

En nuestro ejemplo de los estudiantes de postgrado podemos seleccionar una muestra aleatoria simple de cuarenta elementos utilizando la tabla de números aleatorios. Primero debemos numerar a los estudiantes del 001 al 271. Luego podemos comenzar a seleccionar nuestra muestra leyendo los últimos tres dígitos de cada número aleatorio de la primera fila de la tabla del apéndice D. Pudimos arbitrariamente haber escogido los tres primeros dígitos del número aleatorio y/o avanzar por la primera columna. De igual manera, si el total de estudiantes fuera mayor de 999, tendríamos que haber escogido cuatro dígitos del número aleatorio.

El primer número aleatorio de tres dígitos será 271, el cual corresponde al último estudiante de nuestra lista poblacional, que se convierte en el primer elemento de nuestra muestra. Pasamos al segundo número aleatorio de la primera fila y leemos los tres últimos dígitos, 986. Como no hay ningún estudiante con ese número, pasamos al siguiente número, 744; tampoco existe ningún estudiante con ese número. El siguiente es el 102, y el estudiante con ese número pasa a constituir el segundo elemento de nuestra muestra. Continuamos de esta manera hasta seleccionar a los cuarenta estudiantes para la muestra. Si se encuentra un número que ya fue seleccionado, simplemente se le ignora. Así, los primeros diez estudiantes seleccionados para la muestran serán aquellos a quienes se asignaron los siguientes números:

271, 102, 141, 108, 115, 041, 030, 243, 147, 122

Seleccionar los elementos de una muestra aleatoria simple puede tomar mucho tiempo. Pero la tarea de encontrar y obtener la información acerca de los elementos muestrales puede ser aún más difícil, en especial cuando estos elementos están geográficamente dispersos.

Cuando la población es muy grande y/o heterogénea, resulta muy difícil y costoso conseguir la lista o marco poblacional actualizado de todos los elementos de la población. Por ejemplo, para estudiar la distribución del ingreso familiar en Lima Metropolitana con base en una muestra aleatoria simple, necesitaremos, en principio, una lista o censo poblacional actualizado de todas las familias residentes en Lima Metropolitana, que no está siempre disponible. Por esta razón, el muestreo aleatorio simple no resulta ser el procedimiento más común en la práctica. Sin embargo, es importante prestar atención a este diseño muestral por dos razones. Primero, porque muchos de los diseños muestrales más elaborados usan el muestreo aleatorio simple en algún aspecto de su diseño, y por lo tanto comprender el muestreo aleatorio simple es esencial para entender tales diseños. En segundo lugar, porque muchos procedimientos de muestreo no aleatorios son diseñados para simular el muestreo aleatorio simple en muchos aspectos.

b. Muestra aleatoria sistemática

Muchas veces se puede lograr la aleatoriedad que se obtiene con una muestra aleatoria simple en menos tiempo y con menor esfuerzo, usando una muestra sistemática. Si hay N elementos en la población, enumerados del 1 al N, una muestra sistemática de n elementos es una que comienza en algún elemento entre 1 y N/n escogido aleatoriamente, e incluye cada N/n-ésimo elemento contado a partir del primer elemento seleccionado. Por ejemplo, si se desea obtener una muestra que consista del 1% de una población de 10,000 elementos, entonces seleccionaremos al azar el primer elemento entre los 100 primeros elementos de la

población, y después seleccionaremos del resto de la lista a cada elemento que ocurra a una distancia de 100 elementos con respecto al primero que se selecciona.

Entonces, en una muestra sistemática sólo se necesita seleccionar un número aleatorio para comenzar, y luego se seleccionan los otros elementos a iguales intervalos en la lista de la población. Este procedimiento es más rápido y simple que el muestreo aleatorio simple, y puede ser igual de bueno si es que los elementos de la población están ordenados al azar en el marco poblacional.

c. Muestra aleatoria estratificada

Para seleccionar una muestra aleatoria estratificada se divide la población en grupos o estratos y se seleccionan muestras aleatorias simples en cada uno de ellos. El tamaño de estas muestras es proporcional al tamaño de los estratos correspondientes. El muestreo aleatorio estratificado se justifica cuando tenemos una buena razón para creer que los estratos difieren significativamente unos de otros respecto a la característica que se está midiendo. Luego, la estratificación de la población dependerá del conocimiento que tenga el investigador de las diferencias en los estratos. Por ejemplo, si queremos estudiar el consumo de calorías de las familias de Lima Metropolitana podríamos estratificar a la población por distritos de residencia, bajo la hipótesis de que en las zonas marginales la menor capacidad adquisitiva de las familias se traduce en un menor consumo de calorías.

La estratificación tiene una ventaja adicional, y es que proporciona información tanto para cada estrato separadamente como para toda la población. Pero su costo de obtención es mayor que el de una muestra aleatoria simple del mismo tamaño, ya que la estratificación implica el trabajo adicional de clasificar a la población en los diversos estratos.

B. Muestras no probabilísticas

Si no tenemos una lista de todos los elementos de la población, no será posible obtener una muestra aleatoria ya que no se sabe cuál es la probabilidad de que un elemento de la población sea incluido en la muestra. Cuando no se dispone del marco poblacional, la inclusión o no de un elemento depende muchas veces del criterio del investigador. El problema de confiar en el criterio de una persona para seleccionar una muestra, aun cuando dicha persona tenga un buen conocimiento del caso, es que no se puede medir el error de muestreo, el cual sí puede calcularse cuando la muestra es tomada con base en probabilidades. Esto quiere decir que en las muestras no probabilísticas no es posible determinar la probabilidad de inclusión de cada elemento de la población en la muestra y que, por tanto, no hay forma de medir el riesgo de llegar a conclusiones erróneas. Dado que la confiabilidad de los resultados de estas muestras no puede medirse, las muestras no probabilísticas no se prestan para el tratamiento y análisis estadístico riguroso. A pesar de estas limitaciones, las muestras no probabilísticas son empleadas como la segunda mejor alternativa cuando no se pueden obtener muestras estrictamente aleatorias. Los tipos más comunes de muestreo no probabilístico son las muestras por cuotas, las muestras por conveniencia y las muestras por criterio.

a. Muestras por cuotas o dirigidas

Es el tipo de muestra no probabilística más usada. El método es especialmente empleado para el diseño muestral de las encuestas de opinión pública. El muestreo por cuotas busca controlar el componente subjetivo en el proceso de selección al especificar exactamente el número de individuos que el encuestador debe seleccionar en cada una de las categorías. Por ejemplo, el encuestador es obligado a obtener información acerca de un número

específico de personas de cierta edad, sexo, nivel de ingreso, etcétera. A pesar de que este método no es aleatorio, al fijar cuotas en las categorías que están más relacionadas a las variables de interés del estudio uno puede obtener resultados que en muchos casos no serán muy inferiores a los de una muestra aleatoria.

La gran ventaja del muestreo por cuotas es su velocidad: no se perderán horas tratando de localizar a las personas que fueron seleccionadas aleatoriamente o reemplazando a aquellas que no desean cooperar. Este muestreo es especialmente útil cuando los resultados deben estar actualizados, como en el caso de las encuestas que buscan medir el estado actual de la opinión pública. Pero el gran problema es que el entrevistador usualmente es capaz de completar sus cuotas más rápidamente yendo a lugares donde la gente está concentrada en pequeñas áreas. De hacerlo así, introducirá sesgos inesperados en la muestra. Por ejemplo, si el entrevistador recurriera a entrevistar personas en un paradero de ómnibus, la muestra así seleccionada no incluirá a personas que no usan servicios de transporte público.

b. Muestras por conveniencia

Los elementos de la población que se incluyen en una muestra por conveniencia se eligen por su facilidad de acceso y conveniencia. Por ejemplo, en un programa de televisión se propone que se deje de construir el tren eléctrico de Lima, y se solicita la reacción del público mediante llamadas por teléfono. No cabe duda de que la gente interesada en el tema llamará, pero de ninguna manera estas personas representarán al total de la población de Lima, aun cuando el número de llamadas sea exageradamente alto. Mucha gente no estará viendo el programa, y de los que ven el programa muchos no tendrán teléfono o simplemente no tendrán interés en el tema.

c. Muestras por criterio

Los elementos de la población que se incluyen en una muestra por criterio son seleccionados por un experto con base en su confianza de que estos elementos son efectivamente representativos de la mayoría de los elementos de la población. Muchas veces se usan estas muestras cuando el costo y/o tiempo disponible hacen que no sea posible tomar una muestra aleatoria, y sólo sea necesaria una muestra pequeña.

3. DISTRIBUCIONES MUESTRALES DE LOS ESTADÍGRAFOS

Después de haber discutido los diferentes tipos de diseño muestral, revisaremos la teoría matemática que nos permite usar una muestra aleatoria simple para estimar las características de la población. La base fundamental de la inferencia estadística es el *teorema del límite central.* Primero discutiremos la aplicación de los resultados de este teorema en la estimación de la media poblacional (μ), y luego en la estimación de la proporción (π).

A. Distribución muestral de la media

Si tomamos repetidamente muestras aleatorias de una población y calculamos la media de la variable X en cada muestra ($\overline{X}_i$), encontraremos que la mayoría de estas medias muestrales difieren una de otra. La distribución de probabilidades asociadas a estas medias muestrales se denomina *distribución muestral teórica de la media.* Así, la distribución muestral de la media incluye todo valor posible que la media muestral puede tomar y las probabilidades correspondientes a cada uno de estos valores. Esta distribución muestral de la media tiene un valor esperado, denotado por el símbolo $\mu_{\overline{x}}$, y una desviación estándar o *error estándar*, denotado por $\sigma_{\overline{x}}$.

Hay dos teoremas importantes que relacionan la distribución muestral de la media con las características de la población original.

– *Teorema 1.* Si tomamos repetidamente muestras aleatorias de tamaño n de una población, las características de la distribución muestral de la media estarán dadas por:

$$\mu_{\bar{x}} = E(\bar{X}) = \mu \qquad (1)$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (2.a)$$

o, si n 0.05 N,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{n-1}} \qquad (2.b)$$

donde μ y σ son los parámetros de la población de donde se tomaron las muestras de tamaño n.

La ecuación (1) expresa que el valor esperado de la distribución de las medias muestrales, $\mu_{\bar{x}} = E(\bar{X})$, es igual a la media de la población original, μ. Nótese que para que esto sea cierto, debemos tomar todas las muestras posibles de tamaño n de la población finita; si tratamos con una población infinita, debemos continuar tomando muestras aleatorias de tamaño n indefinidamente.

La ecuación (2.a) establece que el error estándar de la media, $\sigma_{\bar{x}}$, está dado por la desviación estándar de la población, σ, dividido por la raíz cuadrada del tamaño de la muestra, n. Para poblaciones finitas de tamaño N debe introducirse un *factor de corrección por población finita,* $f = \sqrt{(N-n)/(N-1)}$; entonces, la fórmula de $\sigma_{\bar{x}}$ está dada por la ecuación (2.b). Sin embargo, si el tamaño de la muestra es muy pequeño en relación al tamaño de

la población, f se aproxima a 1 y puede eliminarse de la fórmula. Por convención se usará la ecuación (2.b) siempre y cuando $n > 0.05\ N$.

– *Teorema* 2. Teorema del límite central: A medida que el tamaño de la muestra se incrementa ($n \rightarrow \infty$), la distribución muestral de la media se aproxima a la distribución normal, independientemente de la distribución de la población original de la cual se obtuvieron las muestras. La aproximación es suficientemente buena para $n > 30$.

Simbólicamente podemos resumir el teorema del límite central de la siguiente manera:

$$\overline{X} \sim N(\mu, \sigma_{\overline{x}}) \tag{5.3}$$

Esta expresión nos dice que las medias muestrales ($\overline{X}$) se distribuyen (~) normalmente (N) con una media igual a μ y un error estándar igual a $\sigma_{\overline{X}}$.

De las ecuaciones (2.a) y (2.b) vemos que la dispersión de las medias muestrales, medida por $\sigma_{\overline{X}}$, está directamente relacionada a σ, e inversamente relacionada a $\sqrt{n}$. Así, al incrementar el tamaño de la muestra en cuatro veces, se incrementa la exactitud de X como una estimación de μ reduciendo $\sigma_{\overline{X}}$ a la mitad. Nótese también que $\sigma_{\overline{X}}$ es siempre más pequeño que σ. La razón para esto es que las medias muestrales, como promedios de las observaciones de las muestras, presentan menos variabilidad o dispersión que los valores poblacionales. Aun más: a medida que se incrementa el tamaño de la muestra, mayores serán las reducciones de los valores de $\sigma_{\overline{X}}$ con relación al valor de σ.

Debemos resaltar que cuanto más pequeña sea $\sigma_{\overline{X}}$, más exacta será la media muestral $\overline{X}$ como una estimación de la media poblacional μ. Por está razón $\sigma_{\overline{X}}$ se denomina usualmente *error estándar* de la media.

Dado el teorema del límite central, el estadígrafo z, definido por:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad (4)$$

se aproxima a la distribución normal estandarizada a medida que $\bar{X}$ se incrementa. Luego, podemos calcular la probabilidad de que la media $\bar{X}$ de una muestra aleatoria se encuentre en un intervalo específico localizando los valores correspondientes de la variable transformada z en la tabla del apéndice B, como se explicó en la sección 7 del capítulo III.

Consideremos el siguiente ejemplo para ilustrar el uso del teorema del límite central. Suponga que la distribución de probabilidades de ingresos familiares mensuales en la ciudad de Arequipa es asimétrica positiva ($\mu > M_0$), con un valor esperado de \$ 500 y una desviación estándar de \$ 700. Si se toma una muestra aleatoria de cien familias, ¿cuál es la distribución muestral de los ingresos familiares promedio de muestras de tamaño n = 100?

De acuerdo con el teorema del límite central, $\bar{X}$ tendrá una distribución normal de probabilidades con:

$$E(\bar{X}) = \mu = \$500$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{700}{\sqrt{100}} = \$70$$

Con estos parámetros podemos determinar la probabilidad de que $\bar{X}$ se encuentre dentro de cualquier intervalo. Por ejemplo, que el ingreso promedio muestral sea mayor que \$600, menor que \$ 300, ó que se encuentre entre \$400 y \$700. Esto será posible calculando los valores correspondientes de z y determi-

nando las probabilidades acumuladas en las tablas de la distribución estandarizada. Así:

$$p(\overline{X} > 600) = P[z > (600 - 500)/70] = p(z > 1.43) = 0.076$$

La probabilidad de que el promedio de ingresos familiares mensuales de una muestra de cien familias sea mayor que $ 600 es de 7.6%.

B. Distribución muestral de la proporción (p)

De manera similar, los resultados del teorema del límite central se aplican al muestreo para estimar una proporción poblacional. Primero, definimos la *distribución muestral teórica de la proporción* (p) como la distribución de probabilidades de la proporción de éxitos en n pruebas independientes (i.e., $p = x/n$, donde x es el número de éxitos) de una población con una probabilidad de éxito de π. Luego, esta distribución muestral de la proporción tiene una media o valor esperado, μ_p, y una desviación estándar o *error estándar* de la proporción, σ_p.

- *Teorema 1'*. Si tomamos repetidamente muestras aleatorias de tamaño n de una población con una probabilidad de éxito de π, las características de la distribución muestral de la proporción estarán dadas por:

$$\mu_p = E(p) = \pi \tag{5}$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \tag{6.a}$$

o, si $n > 0.05\,N$:

$$\sigma_p \sqrt{\frac{\pi(1-\pi)}{n}}\sqrt{\frac{N-n}{N-1}} \tag{6.b}$$

Análogamente, tenemos el teorema del límite central aplicado a la proporción.

– *Teorema 2'*. A medida que el tamaño de la muestra se incrementa ($n \to \infty$), la distribución muestral de la proporción se aproxima a la distribución normal.

$$p \sim N(\pi, \sigma_p) \qquad (7)$$

Para ilustrar los resultados de estos teoremas supongamos que el 30% de la población de Lima está de acuerdo con la propuesta de construir un tren eléctrico. ¿Cuál es la distribución muestral de la proporción de gente que favorece la propuesta en una muestra aleatoria de cien personas? ¿Cuál es la probabilidad de que más del 50% de las personas incluidas en la muestra esté de acuerdo con la propuesta?

La distribución muestral de la proporción de personas a favor de la construcción del tren eléctrico, p, será una distribución de probabilidades normal, ya que $n > 30$ con:

$$\mu_p = E(p) = \pi = 0.30$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(0.3)(0.7)}{100}} = 0.046$$

La probabilidad de que p sea mayor que 0.50 estará dada por:

$$p(p > 0.50) = p\left(z > \frac{0.500 - 0.300}{0.046}\right) = p(z > 4.46)$$

$$= 0.500 - 0.499 = 0.001$$

La probabilidad de que más del 50% de las personas incluidas en una muestra de cien estén de acuerdo con la propuesta de construir el tren eléctrico es 0.001.

4. ESTIMACIÓN DE PARÁMETROS

Una muestra aleatoria de treinta alumnos de pregrado de la Universidad del Pacífico obtuvo 13 como promedio de notas durante el último ciclo de estudio. ¿Qué nos dice este resultado con respecto a la nota promedio del total de los 1,326 alumnos de pregrado? Dos conclusiones extremas e igualmente falsas se podrían derivar de esta información. Una es la conclusión inocente según la cual la nota promedio de todos los alumnos es exactamente 13. Esta conclusión se basa en el supuesto de que lo que es cierto para la muestra es cierto para la población. En el otro extremo tenemos la posición escéptica, que arguye que no podemos concluir nada de los datos, dado que una muestra de treinta es insignificante comparada con los restantes 1,296 estudiantes que no fueron incluidos en la muestra.

Ambas conclusiones, una basada en una fe ciega en las muestras y la otra en un escepticismo total, son inválidas. La fe no es justificada; el escepticismo no es necesario. De la sección anterior sabemos que las medias muestrales están sujetas a variaciones probabilísticas, y que ninguna de ellas puede aceptarse como la media poblacional. Sin embargo, también se señaló que tal variación puede ser medida, permitiéndonos formular aseveraciones acerca de la población con un cierto grado de confianza. Esto nos lleva a introducir el concepto de *intervalo de confianza*.

Es posible obtener una estimación *puntual* o una estimación por *intervalo* de un parámetro de la población. Como dijimos anteriormente, por razones de costo, tiempo y viabilidad los parámetros de la población son frecuentemente estimados sobre la base de estadígrafos muestrales. El estadígrafo muestral usado para estimar un parámetro de la población se llama *estimador*, y un valor observado específico se denomina *estimación*. Cuando la estimación de un parámetro de la población está dada solamente por un número, se denomina estimación puntual. Por ejemplo, la media muestral $\overline{X}$ es un estimador de la media de la

población, μ, y un valor simple de $\overline{X}$ es una estimación puntual de μ. Del mismo modo, la proporción de la muestra, p, se puede usar como estimador para la proporción de la población, π, y un valor simple de p es una estimación puntual de π.

Una estimación puntual es no-sesgada si en muestras aleatorias repetidas de la población el valor esperado del estadígrafo correspondiente es igual al parámetro de la población que se estima. Por ejemplo, $\overline{X}$ es una estimación puntual no-sesgada de μ porque, como lo muestra la ecuación (1) de este capítulo, $E(\overline{X}) = \mu$. De igual manera, la proporción de la muestra, p, es una estimación no-sesgada de π.

Una *estimación por intervalo* de un parámetro de la población consiste en definir un rango de valores dentro del cual esperamos se encuentre dicho parámetro con una probabilidad preestablecida llamada *nivel de confianza*. Este rango de valores se conoce como *intervalo de confianza*, y está usualmente centrado alrededor de la estimación puntual no-sesgada. Si suponemos que la desviación estándar de la población (σ) es conocida y disponemos de una muestra aleatoria con más de treinta elementos, podemos encontrar el intervalo dentro del cual se encontrará la media poblacional (μ)con un nivel de confianza del 95%.

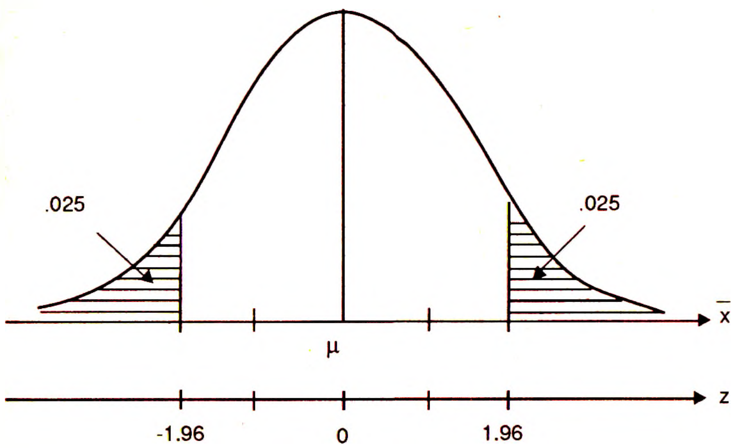
$$p\,(\overline{X} - 1.96\,\sigma_{\bar{x}} < \mu < \overline{X} + 1.96\,\sigma_{\bar{x}} = 0.95) \qquad (8)$$

donde 1.96 es el valor de la variable normal estandarizada asociado a este nivel de confianza, como se muestra en el gráfico 34.

La ecuación (8) establece que en un muestreo aleatorio repetitivo esperamos que el 95% de los intervalos construidos alrededor de las medias muestrales, $\overline{X}$, incluya a la media poblacional. Luego, podemos afirmar que existe una probabilidad del 95% de que un intervalo de confianza específico, basado en una muestra aleatoria simple, incluya el valor de la media poblacio-

nal. Los dos valores que definen el intervalo de confianza ($X \pm 1.96\,\sigma_{\bar{X}}$) se denominan *límites de confianza.* Dado que una estimación por intervalo también expresa el grado de exactitud o confianza que tenemos en la estimación, se la prefiere a la estimación puntual.

Gráfico 34: Distribución normal para determinar el valor de z necesario para una confianza del 95%



Debemos señalar que el concepto de intervalo de confianza está relacionado al resultado del teorema del límite central, que establece que $\bar{X} \sim N(\mu, \sigma_{\bar{X}})$. Este teorema nos permite definir que:

$$p\,(\mu - 1.96\,\sigma_{\bar{x}} < \bar{X} < \mu + 1.96\,\sigma_{\bar{X}}) = 0.95 \qquad (9)$$

es decir, que la probabilidad de que una media muestral se encuentre en el intervalo $[\mu - 1.96\,\sigma_{\bar{x}},\ \mu + 1.96\,\sigma_{\bar{x}}]$ es del

95%. Por otro lado, la ecuación (8) nos define el siguiente intervalo:

$$[\overline{X} - 1.96\ \sigma/\sqrt{n},\ \overline{X} + 1.96\ \sigma/\sqrt{n}] \qquad (10)$$

que tiene una amplitud de 2 * (1.96 $\sigma_{\overline{x}}$), exactamente como el intervalo de la ecuación (9). La única diferencia es que la ecuación (8) centra el intervalo en torno de $\overline{X}$, y no en torno de μ, como establece el teorema del límite central. Como la probabilidad de obtener un valor de $\overline{X}$ en el intervalo $\mu \pm 1.96\ \sigma_{\overline{x}}$ es de 0.95, se deduce que el 95% de todos los intervalos generados por la ecuación (10) incluirá μ.

Podemos concluir diciendo que una vez obtenida la muestra y calculado el intervalo dado por la ecuación (10), hay una confianza del 95% de que la media poblacional esté contenida en este intervalo.

Para ilustrar el concepto del intervalo de confianza, consideremos el siguiente ejemplo: Se tomó una muestra aleatoria de 121 estudiantes de los 1,326 alumnos de pregrado y se encontró que la calificación promedio era 13. De estudios previos se sabe que la desviación estándar de las notas de la población estudiantil es 5.2. El intervalo de confianza al 95% para la calificación promedio de la población total de estudiantes es:

$$[\overline{X} \pm 1.96\ \sigma_{\overline{x}}]$$

Puesto que $n > 0.05\ N$, reemplazamos el valor de $\sigma_{\overline{x}}$:

$$[\overline{X} \pm 1.96\ \frac{\sigma}{\sqrt{n}}\ \sqrt{\frac{N-n}{N-1}}]$$

$$[13 \pm 1.96\ \frac{5.2}{\sqrt{121}}\ \sqrt{\frac{1,326-121}{1,326-1}}]$$

$$[13 \pm 1.96\,(0.47)\,(0.95)]$$

$$[13 \pm 0.88]$$

Luego, diremos que la calificación promedio de los estudiantes de pregrado, μ, estará entre 12.12 y 13.88, con un nivel de confianza del 95%.

A. Intervalo de confianza para la proporción

De manera similar, se puede construir un intervalo de confianza para la proporción de la población. Usando las ecuaciones (5) a (7), y dado que el tamaño de la muestra aleatoria es igual o mayor que 30, podemos encontrar el intervalo de confianza del 95% para la proporción de la población, π, como:

$$[p - 1.96\,\sigma_p,\ p + 1.96\,\sigma p] \qquad (10')$$

Otros intervalos usados con frecuencia son aquellos con niveles de confianza del 90 y 99%, correspondientes a los valores de z de 1.64 y 2.58, respectivamente (ver la tabla del apéndice B).

B. Intervalo de confianza cuando σ es desconocido

El intervalo de confianza definido por la ecuación (10) se calcula tomando como base un valor conocido de σ. En muchos casos, en los que la población no ha sido estudiada anteriormente, no se conocerá este parámetro. Sin embargo, será posible estimarlo con base en la desviación estándar, s, de los elementos de una muestra de tamaño suficientemente grande. El estimador nosesgado de σ, para muestras de treinta o más elementos, está definido por:

$$s = \sqrt{\frac{\Sigma\,(X - \overline{X})^2}{(n-1)}} \qquad (11)$$

Este estimador se usará para calcular el error estándar de la media:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{12}$$

entonces, el intervalo de confianza del 95% será:

$$[\bar{X} + 1.91\ s/\sqrt{n},\ \bar{X} + 1.96\ s/\sqrt{n}] \tag{13}$$

De manera similar, el intervalo de confianza para la proporción definido por la ecuación (10') requiere conocer el error estándar de p, (σ_p), cuyo valor se calcula como $\sqrt{n(1-\pi)}$. Puesto que π es desconocido, podemos usar la proporción muestral p como un estimador de π, siempre que se tenga una muestra de tamaño suficientemente grande ($n \geq 30$).

Es decir, cuando n es grande, el intervalo de confianza del 95% para la proporción de la población, π, se puede definir como:

$$[p - 1.96\ s_p,\ p + 1.96\ s_p] \tag{14}$$

donde

$$s_p = \sqrt{\frac{p\,(1-p)}{n}} \tag{15}$$

5. INTERVALOS DE CONFIANZA USANDO LA DISTRIBUCIÓN T

Cuando la población bajo estudio tiene una distribución normal (o aproximadamente normal) pero no se conoce la desviación estándar de la población, y el tamaño de la muestra es menor que 30, no podemos recurrir al teorema del límite central; es decir, no podemos usar la distribución normal para determinar intervalos de confianza para la media de la población. En estos casos utilizaremos la distribución t de Student.

Al igual que la distribución normal estandarizada, la distribución t tiene forma de campana y es simétrica alrededor de la media cero, pero es más «achatada» que la distribución normal estandarizada debido a su mayor dispersión. Es decir, la variación de las medias muestrales de muestras pequeñas es mayor que aquella de muestras grandes. Esta dispersión será mayor porque, si tomamos una muestra pequeña ($n < 30$), se podría seleccionar un elemento que se encuentre en uno de los extremos, cola superior o inferior, y habrá poca probabilidad de que este valor sea compensado por otro elemento de la muestra que se ubique en el otro extremo de la distribución poblacional. Consecuentemente, la media muestral calculada podría estar distante de la media poblacional.

Mientras que la distribución normal estándar es una sola, la distribución t es diferente para cada tamaño de muestra. El estadístico t se define como:

$$t_{(n-1)} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \qquad (16)$$

La tabla del apéndice C muestra los valores de t para diferentes porcentajes de área bajo la curva de la distribución de Student y diferentes *grados de libertad* (gl). Los grados de libertad están especificados en la primera columna. El área total bajo la curva es igual a 1. Los valores de la primera fila (0.10, 0.05, 0.025,...) miden la proporción del área bajo la curva a la derecha del valor de t. Así, por ejemplo, para el área sombreada que representa el 0.05 del área total bajo la curva, el valor de t para 20 grados de libertad es 1.725. Los grados de libertad se refieren al número de valores que pueden tomarse libremente. Por ejemplo, si trabajamos con una muestra de dos elementos y sabemos que la media muestral para estos dos valores es 20, sólo podemos asignar libremente el valor de uno de estos dos elementos. Supongamos que asignamos libremente un valor de 24 a uno de

esos elementos; el otro debe ser necesariamente igual a 16 para obtener una media de 20. Entonces decimos que tenemos (n - 1) = 2-1 = 1 grado de libertad. Igualmente, n = 15 significa que podemos asignar libremente valores a catorce de estos quince elementos si queremos estimar la media poblacional, y así tendremos (n - 1) = 14 grados de libertad. En conclusión, para poder usar la distribución t de Student los grados de libertad se definen como el tamaño de la muestra menos uno [gl = (n - 1)].

El intervalo de confianza del 95% para la media poblacional usando la distribución t, está dado por:

$$[\overline{X} - t_{(n-1)} \frac{s}{\sqrt{n}} , \overline{X} + t_{(n-1)} \frac{s}{\sqrt{n}}] \quad (17)$$

donde $t_{(n-1)}$ se refiere al valor de t de la distribución de Student, tal que el 2.5% del área total bajo la curva esté ubicado en cada una de las dos colas, a la derecha de t y a la izquierda de -t con (n-1) grados de libertad. Utilizamos $s/\sqrt{n}$ como el estimador de la desviación estándar de las medias muestrales, $s_{\overline{X}}$, en lugar de $\sigma/\sqrt{n}$, ya que no se conoce σ.

Debemos señalar que a medida que n aumenta, la distribución t se aproxima a la distribución normal estándar:

$$t_{(n-1)} \rightarrow z \text{ , cuando } n \rightarrow \infty$$

Como regla práctica, podemos decir que cuando n > 30, las distribuciones t y z son aproximadamente iguales.

6. TAMAÑO MUESTRAL

El concepto de intervalo de confianza nos permite contestar una de las preguntas más comunes entre quienes emplean el muestreo para estimar las características de una población: ¿cuál debe ser el tamaño de la muestra? Si observamos las ecuaciones (10)

y (14), que definen los intervalos de confianza, podemos ver cuán importante es el tamaño muestral para reducir la amplitud de estos intervalos. Un mayor tamaño muestral proporciona un intervalo de confianza de menor amplitud, y la respuesta natural a la pregunta respecto al tamaño muestral es que entre más grande sea la muestra, mejor. Pero esta no es una respuesta práctica si el muestreo es costoso y/o toma tiempo, o si al tomar la muestra se destruye el producto que se está probando. Para determinar el tamaño muestral debemos contestar a otra pregunta: Dado el costo del muestreo, ¿cuál es la magnitud máxima del "error permisible" que estamos dispuestos a aceptar?

En el caso de la estimación de la media poblacional, el valor de $1.96 * (\sigma/\sqrt{n})$ se suma y se resta de $\overline{X}$ para determinar los límites del intervalo de confianza del 95%. En otras palabras, podemos decir que la diferencia entre la media poblacional, μ, y la media muestral, $\overline{X}$, es menor o igual a $1.96 * (\sigma/\sqrt{n})$ con una probabilidad del 95%. Esta diferencia será el error tolerable:

$$E = 1.96\,(\sigma/\sqrt{n}) \qquad (18)$$

En otras palabras, E equivale a la mitad de la amplitud del intervalo de confianza.

Para determinar el tamaño muestral óptimo debemos establecer el tamaño del error permisible, E, que estamos dispuestos a aceptar, considerando que mantener un E muy pequeño puede conducirnos a un tamaño muestral muy grande. Despejando n de la ecuación (18), tenemos:

$$n = \frac{(1.96)^2\,\sigma^2}{E^2} \qquad (19)$$

y en general, para diferentes niveles de confianza, el tamaño óptimo de n estará dado por:

$$n = \frac{z^2 \sigma^2}{E^2} \qquad (20)$$

Es decir, el tamaño muestral depende del grado de confianza deseado reflejado por el valor de z, de la variabilidad de la población de la cual se obtiene la muestra (σ^2) y del error permisible (E). Observando la ecuación (20) podemos decir que el tamaño muestral debe aumentarse si: a) se fija un error permisible menor; b) el grado de confianza (z) deseado aumenta; o, c) si la variabilidad de la población aumenta.

La ecuación (20), que define el tamaño muestral óptimo, supone que el valor de σ^2 es conocido. En muchos casos en los que no conocemos μ, tampoco conoceremos σ. Una alternativa podría ser utilizar el conocimiento que se tiene de la varianza de poblaciones similares a la de nuestro interés. Una segunda alternativa es tomar una muestra piloto de tamaño n_0. Con los datos de esta pequeña muestra preliminar podemos calcular la desviación estándar de la muestra, s, y usar la distribución t para estimar el tamaño muestral óptimo:

$$n = \frac{t^2 s^2}{E^2} \qquad (21)$$

Sobre la base de este valor de n, determinamos cuántos elementos más debemos seleccionar de la población, $(n - n_0)$, para obtener el tamaño muestral óptimo.

Ahora, supongamos que deseamos determinar el tamaño muestral óptimo para estimar la proporción con un nivel de confianza del 95%. Al igual que en la estimación de μ, necesitamos decidir el tamaño del error permisible, $E = 1.96 * \sqrt{\pi(1 - \pi)}/\sqrt{n}$), que estamos dispuestos a aceptar. Entonces el tamaño muestral óptimo estará dado por:

$$n = \frac{(1.96)^2 \pi(1-\pi)}{E^2} \tag{22}$$

y en general, si no fijamos el nivel de confianza en 95%, el tamaño óptimo de la muestra será:

$$n = \frac{z^2 \pi (1-\pi)}{E^2} \tag{23}$$

Para usar esta fórmula necesitamos el valor de π, pero es precisamente este parámetro el que tratamos de estimar mediante la muestra. Hay varias posibilidades para obtener un estimado preliminar de π. Al igual que el caso de σ, una primera alternativa es usar información de una variable similar para la cual se conoce π. Una segunda posibilidad es tomar una muestra piloto de tamaño n_o, y usar la proporción muestral, p, como un estimado de π. Usando la distribución t para estimar el tamaño muestral óptimo, tenemos:

$$n = \frac{t^2 p (1-p)}{E^2} \tag{24}$$

Luego completamos la muestra tomando $(n - n_o)$ elementos adicionales de la población, para obtener el tamaño muestral óptimo.

Un tercer método consiste en suponer que π es 0.5. Este método, aparentemente simplista, es el más conservador, pues exige el valor máximo de n, para valores fijados de z y E. Luego, en ausencia de toda información acerca del posible valor de π, podemos asignarle el valor de 0.5 con la seguridad de que el tamaño muestral así obtenido será suficiente para satisfacer los requisitos establecidos respecto al nivel de confianza y al error permisible, independientemente del valor exacto de π. Final-

mente, debemos señalar que el valor de n, definido por la ecuación (23), no es muy sensible a variaciones del valor de π alrededor de 0.5, siempre que π esté entre 0.3 y 0.7. Consecuentemente, los errores en un estimador original de π no tendrán grandes efectos en el tamaño de n.

Para ilustrar este último método consideremos el siguiente ejemplo: Se quiere estimar, con un nivel de confianza del 90%, la proporción de ciudadanos de Lima Metropolitana que votarían por el candidato del Partido Liberal con un error permisible de 0.05. ¿Cuál es el tamaño óptimo de la muestra requerida?

Para usar la ecuación (23) necesitamos un estimado de π, la proporción poblacional que favorece al candidato del Partido Liberal. Como no conocemos π, podemos suponer que es igual a 0.5, lo que nos asegura que no subestimamos el tamaño de la muestra. Entonces:

$$n = \frac{(1.64)^2\,(0.5)\,(0.5)}{(0.05)^2} = 269$$

Una muestra de 269 personas nos permitirá obtener un estimador no-sesgado de la proporción de ciudadanos que votarían por el candidato del Partido Liberal.

● *Ejercicios*

1. Defina y discuta los siguientes conceptos:

a. Intervalo de confianza.
b. Muestras aleatorias y no aleatorias.
c. Distribución muestral de la media.
d. Error tolerable.

2. La tienda Wong calculó el valor promedio de las compras de 49 clientes que adquirieron sus artículos usando tarjeta de crédito. La media resultó \$ 60.50, y la desviación estándar \$ 20.10.

Encuentre un intervalo de confianza del 98% para el valor promedio de compras de todos los clientes que usan tarjetas de crédito.

3. Se extrajo una muestra aleatoria de 100 familias en la ciudad de Chiclayo para determinar su consumo mensual de electricidad; se anotó el consumo de electricidad durante el mes de setiembre. El total del consumo de electricidad de las cien familias fue 36,000 kilovatios/hora.

$$(\sum x = 36{,}000 \text{ y } \sum x^2 = 16'000{,}000)$$

a. Estime el consumo mensual promedio de electricidad para las familias de Chiclayo.
b. Calcule el error estándar de este estadígrafo.
c. Construya el intervalo de confianza del 95% para el consumo mensual promedio de las familias de Chiclayo.

4. El barrio de Juan Pérez está constituido por 900 familias. Se tomó una muestra de 100 familias y se encontró que 54 de ellas ven regularmente el noticiario "24 Horas" por televisión. Calcule un intervalo de confianza del 95% para el total de familias que sintonizan regularmente "24 Horas". Use el factor de corrección para poblaciones finitas, ya que $n/N = 0.11$.

5. Los estudios de factibilidad de proyectos requieren de una medida de demanda para determinar la rentabilidad de dicho proyecto. En un estudio para determinar la factibilidad de aumentar la programación en el canal de televisión del Estado entre las 11:00 p.m. y las 2:00 a.m., un investigador tomó una muestra aleatoria de 120 viviendas con televisor. Encontró que 51 de estos hogares tenían el televisor encendido en dichas horas por lo menos dos veces a la semana. Encuentre un intervalo de confianza del 92% para la proporción de hogares con televisor cuyos miembros ven televisión entre las 11:00 p.m. y las 2:00 a.m. por lo menos dos veces a la semana.

6. Se extrajo una muestra de 330 empleados del total de los 20,000 trabajadores de "Gas Perú". Se encontró que 139 de los trabajadores en la muestra poseían automóvil. Con base en los datos de la muestra, estime el porcentaje de los trabajadores de esa empresa que poseen automóvil, y establezca un intervalo de confianza de 95% para su estimado del porcentaje del universo.

7. El Departamento de Investigación de Mercado de una compañía fabricante de detergentes realizó una encuesta para averiguar qué proporción de las amas de casa prefiere su marca "Magia Gris". Cincuenta de 72 amas de casa prefieren "Magia Gris". Si las 72 amas de casa representan una muestra aleatoria de la población de todos los compradores potenciales, estime la proporción del total de amas de casa que prefieren "Magia Gris". Use un intervalo de confianza del 92%.

8. La cadena de tiendas "Todos" tiene 58 establecimientos en la ciudad de Lima. El Gerente reúne los datos de ventas diarias de cinco tiendas escogidas al azar: 18, 24, 22, 26 y 16 miles de dólares respectivamente.

a. Calcule las ventas medias diarias y la correspondiente desviación estándar para la muestra.

b. Estime la desviación estándar de la población.

c. El Gerente quiere tener un estimado del verdadero promedio de ventas diarias de todas las tiendas usando un intervalo de confianza del 98%.

9. Se extrae una muestra aleatoria de diez bolsas de leche en polvo de un lote que fue envasado por una cierta máquina. Se encontraron los siguientes pesos netos: 173, 180, 185, 187, 171, 184, 175, 186, 180 y 179.

a. Estime la media de los pesos de las bolsas envasadas por esta máquina.

b. Establezca un intervalo de confianza del 95% para el peso promedio de las bolsas envasadas por esta máquina.

10. Una empresa que vende artefactos electrodomésticos extrajo una muestra al azar de 10 de sus 98 vendedores para determinar el valor de las ventas semanales efectuadas por cada uno de ellos. Los resultados fueron los siguientes: 2,100, 2,200, 1,600, 2,100, 2,400, 2,100, 2,300, 2,200, 1,900 y 1,800 dólares respectivamente.

a. Estime la media de las ventas semanales por vendedor para esta compañía.

b. Calcule el error estándar de este estadígrafo.

c. El Gerente de la empresa desea obtener un estimado del promedio de las ventas semanales de todos sus vendedores con un nivel de confianza del 97%.

11. Se le ha pedido que calcule el porcentaje de familias que planea viajar a provincias en las próximas Fiestas Patrias. Usted no tiene idea de cuál pudiera ser ese porcentaje. Además, se le pide que usted esté razonablemente seguro de que su estimado no diferirá del valor verdadero en más de 6%. Calcule el tamaño necesario para una muestra que tenga esta precisión.

12. Al medir el tiempo de reacción, un psicólogo estima que la desviación estándar de esta medida es de 0.05 segundos. ¿Qué tamaño de muestra debe tomar para tener el 95% de confianza de que el error del estimado del tiempo de reacción promedio no exceda 0.01 segundos?

13. Un administrador universitario desea estimar la proporción de estudiantes matriculados en programas de posgrado en administración de empresas en la ciudad de Lima que tienen estudios de pregrado en

contabilidad dentro de ± 0.05 con 90% de confianza. ¿Qué tamaño de muestra debe estudiar como mínimo, si no hay una base para estimar el valor aproximado de la proporción antes de tomar la muestra?

14. Un equipo de investigación médica se siente seguro del desarrollo de un suero, el cual curará a cerca del 80% de los pacientes que sufren de cierta enfermedad. ¿Cuán grande debe ser la muestra para que el grupo pueda estar seguro en un 98% de que la proporción muestral de los pacientes que se curan esté dentro de más o menos 0.04 de la proporción de todos los casos que el suero curará?

15. La tienda "Artículos Eléctricos" desea utilizar una muestra aleatoria para determinar la vida promedio de los focos que ha recibido en un embarque que consta de 20,000 focos y usted ha sido comisionado para determinar el tamaño de la muestra. La empresa desea tener un nivel de confianza de 98% en la precisión del estimado y que este no fluctúe en más de diez horas, por encima o por debajo del verdadero valor de vida promedio de los focos. Se sabe, además, con base en una muestra que se extrajo de un embarque previo, que la desviación estándar de la población es de 77 horas en la vida de los focos.

16. El Director de una escuela de postgrado que tiene 1,000 estudiantes en total desea estimar el porcentaje de estudiantes que preferirían que el método de enseñanza fuera a través de "estudio de casos" en lugar del método tradicional de "clases magistrales". Calcule el tamaño muestral necesario para esta estimación, si se requiere un nivel de confianza de 98% y un error tolerable no mayor que ± 5%. ¿Qué pasa si se reduce el nivel de confianza a 95%?

17. El Contador de la tienda "SEGO" ha observado que el total de cuentas pendientes por tarjetas de crédito de la propia firma ha aumentado a un nivel alarmante. Para formular una política de descuento por pago oportuno, se necesita conocer el tamaño o magnitud de las cuentas retrasadas.

Suponga que se desea estimar la cantidad total en dólares de las cuentas pendientes que tengan por lo menos noventa días de atraso. ¿Cuál debe ser el tamaño de la muestra si se fija el error tolerable en $ 10,000, con un nivel de significación de 0.10? Suponga que hay 1,500 cuentas distintas con noventa o más días de atraso, y que se piensa que la varianza de estas cuentas es inferior a 5,000.

18. La cafetería de la universidad quiere estimar el grosor promedio del queso que rebana una máquina. Se sabe, por experiencia pasada, que la desviación estándar es de 0.05 cm. ¿Cuántas rebanadas se deben chequear para tener un nivel de confianza del 95% de que el grosor promedio esté entre ± 0.013 cm?

V. Prueba de hipótesis

1. Tipos de errores en la prueba de hipótesis. 2. El teorema del límite central y la prueba de hipótesis. 3. Procedimientos para prueba de hipótesis. 4. Medida del error de tipo II. 5. Prueba de hipótesis para la diferencia entre dos medias o dos proporciones.

En el capítulo anterior se aplicó la teoría de probabilidades para estimar ciertas características de la población tomando como base la información de una muestra. Se establecieron estimadores puntuales y por intervalo para la media y proporción de la población. En la discusión se señaló que los estimados estaban sujetos a error, pero que se podían establecer los límites de un intervalo dentro del cual se esperaba encontrar los valores de los parámetros, con cierto nivel de confianza.

En este capítulo continuamos con este mismo tipo de análisis estadístico, pero ahora utilizamos la información muestral para probar hipótesis sobre las características de la población. De hecho, la *prueba de hipótesis* sobre parámetros de la población es el otro aspecto fundamental de la estadística inferencial. En principio, planteamos una hipótesis o supuesto sobre el valor de un parámetro poblacional, el cual es contrastado con el valor del estadígrafo correspondiente calculado con base en una muestra aleatoria. Esta comparación permite aceptar o rechazar la hipótesis original sobre el valor del parámetro. Trataremos de con-

testar a preguntas como: ¿Aceptamos o rechazamos un embarque de bienes? ¿Han cambiado las preferencias electorales en Lima? ¿Aumentaron las ventas como consecuencia de la campaña publicitaria? ¿Qué afirmación podemos hacer acerca de la vida útil media de nuestro producto?

En la primera sección se analizan los dos tipos de errores que se pueden cometer en el proceso de una prueba de hipótesis, y su relación con el concepto de nivel de significación y nivel de confianza de la prueba. La aplicación del teorema del límite central en la prueba de hipótesis se presenta en la sección 2. El procedimiento científico para la prueba de hipótesis se detalla en la sección 3, ilustrándolo con la prueba de hipótesis con respecto a valores de la media de la población y la proporción. La sección 4 describe la manera de calcular el error de aceptar una hipótesis falsa. Finalmente, la sección 5 discute la prueba de hipótesis acerca de la diferencia entre dos medias poblacionales y entre dos proporciones.

1. TIPOS DE ERRORES EN LA PRUEBA DE HIPÓTESIS

En el proceso de probar una hipótesis, el primer paso consiste en hacer una suposición acerca de las características desconocidas de la población. Luego se toma una muestra aleatoria de esa población y, basándonos en la característica muestral correspondiente, aceptamos o rechazamos la suposición o hipótesis de que la población tenga la característica descrita, con un grado particular de confianza. Consideremos el siguiente ejemplo: Hemos ordenado un cargamento de 1,000 sacos de azúcar de 50 kilos cada uno. Al recibir el producto debemos examinar el peso de los sacos para asegurarnos de que tengan el peso correcto. Para esto tomamos una muestra aleatoria de 30 sacos y calculamos su peso promedio, que resulta ser 47.5 kilos. Con base en esta media muestral debemos confirmar o refutar la afirmación del provee-

dor de que el peso promedio de los sacos de azúcar es de 50 kilos; consecuentemente, aceptaremos o rechazaremos el cargamento.

Al usar una muestra para aceptar o rechazar una hipótesis podemos cometer dos tipos de errores. Primero, podríamos *rechazar una hipótesis que es verdadera*, cometiendo el llamado *error de tipo I*. En el ejemplo del cargamento de sacos de azúcar, es posible que el promedio de los 1,000 sacos sea en efecto 50 kilos, pero que se haya seleccionado una muestra que contiene muchos sacos con menos del peso estipulado, y por tanto la media muestral es de 47.5 kilos. Si hubiéramos fijado la regla de rechazar todo cargamento que produzca una media muestral menor a 48 kilos, estaríamos rechazando un cargamento con el peso apropiado. Este error también es conocido como el *riesgo del productor*, debido a que el comprador decide rechazar el cargamento y el productor debe aceptar la mercadería devuelta a pesar de que el cargamento cumple con los requerimientos estipulados.

Existe otra manera de cometer un error en el contraste de hipótesis; esto ocurre cuando se *acepta una hipótesis falsa* y se comete el llamado *error de tipo II*. En el ejemplo de los sacos de azúcar, suponga que tomamos una muestra y encontramos que la media muestral está muy próxima a 50 kilos, lo que nos lleva a aceptar el cargamento. Sin embargo, al embolsar el azúcar en paquetes de dos kilos, nos damos cuenta de que el peso de la mayoría de los sacos estaba por debajo de 50 kilos. A este error también se le llama *riesgo del consumidor*, ya que este acepta un envío que no cumple con las especificaciones requeridas. La distinción entre estos dos tipos de errores se muestra en el cuadro 11.

En el proceso de contraste de hipótesis podemos establecer la probabilidad de cometer tanto el error de tipo I como el error de tipo II. La probabilidad de cometer el error de tipo I es usualmente denotada por la letra griega alfa (α), mientras que la probabilidad de cometer el error de tipo II está representada por

la letra griega beta (β). Cuando la probabilidad de cometer el error de tipo I se hace más pequeña, la probabilidad de cometer el error de tipo II se incrementa. La única manera de reducir tanto α como β es incrementando el tamaño muestral, como veremos en la sección 4.

CUADRO 11: TIPOS DE ERRORES EN LA PRUEBA DE HIPÓTESIS

	Realidad (Estado de la naturaleza)	
Decisión	Hipótesis verdadera	Hipótesis falsa
Aceptar hipótesis	No hay error	Error de tipo II
Rechazar hipótesis	Error de tipo I	No hay error

A la probabilidad de cometer el error de tipo I, rechazar una hipótesis cierta, se le denomina *nivel de significación o significancia*. Los niveles de significación (α) utilizados con frecuencia son de 5% y 1%, aunque también se usan otros valores dependiendo del tipo de problema bajo análisis. El complemento de α, (1 – α), se denomina *nivel de confianza* de la prueba, y por tanto se refiere a la probabilidad de que aceptemos una hipótesis verdadera. Por ejemplo, si utilizamos el nivel de significación al 5%, entonces el nivel de confianza será del 95%. Lo que determina el valor de α es el nivel de riesgo que se desea correr en un problema específico con respecto a cometer el error de rechazar una hipótesis que sea cierta.

2. EL TEOREMA DEL LÍMITE CENTRAL Y LA PRUEBA DE HIPÓTESIS

En el capítulo anterior estudiamos el teorema del límite central, que establece que la distribución muestral de la media se apro-

xima a la distribución normal cuando las muestras son suficientemente grandes ($n \geq 30$), con una media igual a la media poblacional μ, y una desviación estándar definida por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{1.a}$$

o, si $n > 0.05\ N$, la desviación estándar será:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{1.b}$$

Usando estos resultados podemos calcular la probabilidad de que la media, $\overline{X}$, de una muestra aleatoria de tamaño n, se ubique en un intervalo específico. Primero calculamos los valores de la variable normal estandarizada, *z*, asociados a los límites del intervalo usando la transformación conocida:

$$z = \frac{\overline{X} - \mu}{\sigma_{\bar{x}}} \tag{2}$$

Una vez calculados los valores de z, utilizamos la tabla del apéndice B para determinar la probabilidad de que la media muestral calculada se encuentre en el intervalo definido.

El interrogante que debemos resolver ahora es: ¿se puede explicar la diferencia entre el valor de la media muestral y el valor hipotético para la media poblacional, como una variación causada por el muestreo? ¿O en efecto existe una diferencia *significativa* entre ellas? El término significativo se utiliza en el sentido de que exista algún factor que produzca la diferencia observada, en adición a las variaciones causadas por las fluctuaciones aleatorias que ocasiona la extracción de una muestra aleatoria de la población. La respuesta al interrogante planteado se basa en la comparación del valor calculado de z con el valor

de z de la tabla que separa a aquella parte de la curva normal cuya área es igual al nivel de significación establecido.

En el problema de los sacos de azúcar de 50 kilos, supongamos que se conoce la desviación estándar poblacional, $\sigma = 5$ kilos, y decidimos tomar una muestra de 36 sacos; entonces, la desviación estándar de las medias muestrales de 36 sacos será $\sigma_{\bar{X}} = 5/\sqrt{36} = 0.83$ kilos. De acuerdo con el teorema del límite central, la distribución de las medias muestrales de 36 sacos será:

$$\bar{X} \sim N\,(50, 0.83)$$

En el gráfico 35 se representa esta distribución normal. Sabemos que el 68% del área total bajo la distribución normal se encuentra en el rango $[\mu - \sigma_{\bar{X}},\ \mu + \sigma_{\bar{X}}]$; es decir, que el 68% de las medias muestrales de 36 sacos debe ubicarse en el rango [49.17, 50.83]. De igual manera, podemos afirmar que el 95% de las medias muestrales se encontrará entre 48.37 y 51.62 kilos; es decir, en el intervalo $[\mu - 1.96\,\sigma_{\bar{X}},\ \mu + 1.96\,\sigma_{\bar{X}}]$.

Supongamos que el peso promedio de los 36 sacos de la muestra seleccionada anteriormente es 48 kilos. La pregunta a contestar es si la diferencia entre esta media muestral y la media poblacional hipotética es significativa y si, por tanto, se debe rechazar el cargamento. Debemos establecer el nivel de significación (α) antes de tomar la muestra y hacer el análisis, pues siempre existe el peligro de que el valor de z y su probabilidad ejerzan alguna influencia en la elección del nivel de α para rechazar la hipótesis; esto quiere decir que la elección del valor de α a usarse en el proceso de contraste de hipótesis debe ser independiente de los resultados del análisis de la muestra.

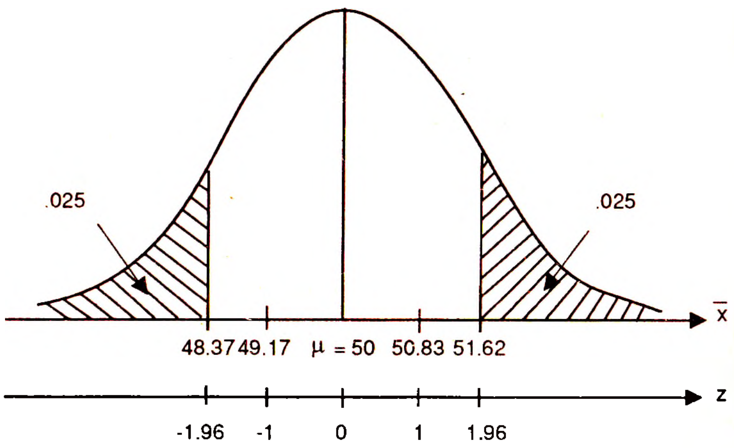
Usando la tabla de la distribución normal estandarizada podemos establecer la probabilidad de que la media de una muestra de 36 sacos extraída de una población que tiene una media de 50 kilos, y una desviación estándar de 5, sea igual o menor a 48 kilos. El gráfico 35 muestra que $\bar{X} = 48$ está en la cola izquierda;

se encuentra a 2.4 errores estándar de μ. De la tabla z podemos determinar que:

$$p(\overline{X} \leq 48) = P(z \leq 2.4) = 0.5 - 0.4918 = 0.0082$$

Existe una probabilidad muy pequeña, 0.82%, de que el cargamento de 1,000 sacos tenga un peso promedio de 50 kilos por saco. Entonces, debemos rechazar el envío y devolver el producto.

Gráfico 35: Distribución de los medios muestrales de 36 sacos de azúcar



Si fijamos $\alpha = 5\%$, la pregunta es: ¿dónde ubicamos el área del 5%? ¿En una o en ambas colas de la distribución? La respuesta depende del punto de vista del decisor. Si el decisor es el recepcionista de sacos del depósito de una cadena de tiendas de comercialización, el área del 5% se ubicará en la cola izquierda,

pues él sólo estará interesado en que el peso promedio del cargamento no sea menor de 50 kilos. Si el peso promedio de los sacos es mayor que 50 kilos, mejor para él. Pero si el decisor es el Jefe de Comercialización de la compañía productora de azúcar, el área del 5% se distribuirá en dos partes iguales en ambas colas, 2.5% en cada una. Esta división se debe a que no será conveniente embolsar sacos con muy poco peso, pues se expone a que los clientes devuelvan los pedidos; ni con mucho peso, pues implicaría pérdidas para la compañía.

Una vez definido el nivel de significación (α) y su distribución en una o dos colas, queda determinada la *región de rechazo o crítica* para el proceso de contraste de hipótesis. Si el valor de una media muestral, $\overline{X}$, cae en el área de rechazo, no estaremos seguros de que la población de la cual proviene la muestra tenga una media de 50 kilos, y por tanto rechazamos el cargamento. Pero si $\overline{X}$ cae en la *región de aceptación*, ($1 - \alpha$), aceptaremos la hipótesis de que el cargamento tiene una media poblacional de 50 kilos.

3. PROCEDIMIENTOS PARA PRUEBA DE HIPÓTESIS

En el proceso de probar una hipótesis acerca de un parámetro de la población con base en la evidencia que proporciona una muestra, calculamos el estadígrafo respectivo de la muestra y decidimos si este valor es razonable para aceptar la hipótesis planteada. Por ejemplo, se desea probar la hipótesis según la cual la mitad de los estudiantes de pregrado de la Universidad del Pacífico son hombres y se toma una muestra aleatoria de 40 estudiantes. Si 22 estudiantes de la muestra son hombres, podríamos, intuitivamente, aceptar la hipótesis planteada, ya que $22/40 = 0.55$ es un resultado muy próximo al valor hipotético de 0.50. Si, por otro lado, 30 de los 40 estudiantes de la muestra fueran hombres, probablemente rechazaríamos la hipótesis original de que la proporción de hombres en la población estudian-

til de pregrado es 0.50, aun a pesar de que este último resultado puede ser posible, de acuerdo con el teorema del límite central. En esta sección discutiremos el procedimiento para la prueba de hipótesis acerca de los diferentes parámetros de la población; asimismo, ilustraremos su aplicación con las pruebas de hipótesis con respecto a la media y a la proporción de la población.

Los pasos formales del procedimiento de prueba de hipótesis son:

(i) Suponer que el parámetro de la población es igual a un valor hipotético. Este valor representa la hipótesis en la cual centramos nuestra atención, conocida como la *hipótesis nula*, y designada por H_0. La *hipótesis alternativa*, representada por H_1, es la afirmación de que el valor del parámetro poblacional es diferente de aquel especificado en H_0.

(ii) Establecer el *nivel de significación*, α, o riesgo de cometer el error de tipo I que estamos dispuestos a aceptar en esta prueba; definir el *tipo de distribución* apropiada de acuerdo con el tamaño muestral; y decidir si se realizará una *prueba de una o dos colas*, dependiendo de la especificación de la hipótesis nula.

(iii) Establecer la *regla de decisión*, una vez definidas la región de aceptación y la región crítica o de rechazo usando la distribución apropiada, z o t.

(iv) Tomar una muestra aleatoria de la población y calcular el estadígrafo respectivo ($\overline{X}$ o p). Si el valor del estadígrafo se ubica en la región de aceptación, se acepta H_0; de lo contrario, se rechaza H_0 en favor de H_1.

Debemos establecer algunas recomendaciones que faciliten la aplicación de este procedimiento en una forma clara y precisa. En primer lugar, todos los posibles valores del parámetro poblacional deben estar incluidos o en la hipótesis nula o en la hipótesis alternativa. En el ejemplo de los estudiantes de pregrado, definiendo π como la proporción de hombres en la población estudiantil, podemos establecer las hipótesis de la siguiente manera:

$$H_0 : \pi = 0.5 \qquad (3)$$
$$H_1 : \pi \neq 0.5$$

Por otro lado, si quisiéramos probar la hipótesis de que existen más hombres que mujeres entre los estudiantes de pregrado, las hipótesis deben establecerse de la siguiente manera:

$$H_0 : \pi \leq 0.5$$
$$H_1 : \pi > 0.5 \qquad (4)$$

En ambos casos, todos los valores posibles de π están siendo cubiertos entre H_0 y H_1. Además, es recomendable que la igualdad del parámetro al valor hipotético esté incluida en H_0.

En segundo lugar, la decisión de usar una prueba de una o dos colas debe basarse en la especificación de la hipótesis nula. Así, si H_0 define una igualdad estricta como en la relación (3), debemos usar una prueba de dos colas, y la región de rechazo estará repartida igualmente en las dos colas de la distribución. En cambio, si H_0 no está definida por una igualdad estricta, como en la relación (4), debemos usar una prueba de una cola, y la región de rechazo se ubicará en uno de los extremos de la distribución. Si $H_0 : \pi \leq 0.5$, la región de rechazo estará en la cola superior.

En tercer lugar, la región de rechazo se define en términos de z o t, pero en algunos casos será recomendable expresarla en términos de las unidades de medida de la media o proporción muestral.

En lo que queda de esta sección ilustraremos la aplicación del procedimiento de prueba de hipótesis con respecto a la media y proporción de la población.

A. Prueba de hipótesis sobre la media de la población

Ejemplo 1: Muestra grande, prueba de dos colas

Supongamos que el decisor en el ejemplo de los sacos de azúcar de 50 kilos es el Jefe del Departamento de Comercialización de

la compañía azucarera. Él desea estar seguro de que los sacos de azúcar que la compañía envía a sus clientes pesan realmente 50 kilos. Se tomó una muestra aleatoria de 36 sacos de azúcar y se encontró que el promedio del peso de esta muestra es 51.5 kilos, con una desviación estándar muestral s = 5 kilos. Si el decisor está dispuesto a aceptar un nivel de significación del 5%, entonces podrá aplicar el procedimiento de prueba de hipótesis como sigue:

(i) $H_0 : \mu = 50$
$H_1 : \mu \neq 50$

(ii) – Nivel de significación: $\alpha = 5\%$.
– Se usará la *distribución z*, puesto que n > 30, y por lo tanto la distribución muestral de la media es normal de acuerdo con el teorema del límite central. Además, podemos usar la desviación estándar muestral, s, como estimación de σ.
– Se tiene una prueba de dos colas; la región de rechazo está en ambos extremos de la distribución, puesto que la hipótesis nula es de una igualdad estricta. Es decir, la región ~~de~~ aceptación de H_0, al nivel de significación del 5%, está dentro de ± 1.96 bajo la curva normal estandarizada (véase el gráfico 36).

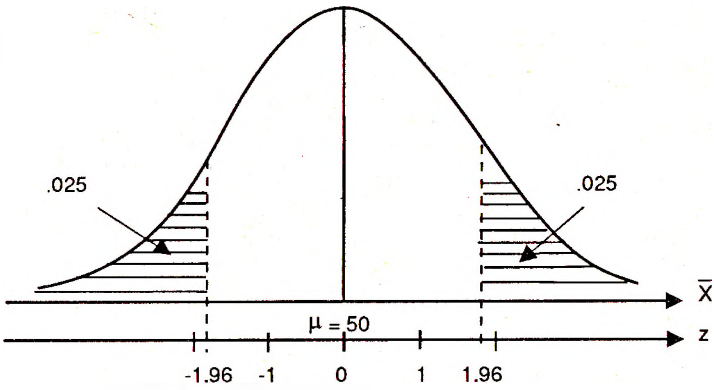
(iii) Regla de decisión. La región de aceptación de H_0, al nivel de significación del 5%, está limitada por los valores de z equivalentes a ±1.96. Consecuentemente, rechazaremos H_0 si $z>1.96$ ó $z<-1.96$, afirmando que el peso promedio de los sacos es diferente de 50 kilos.

(iv) De los datos de la muestra sabemos que $\overline{X}=51.5$ kilos. Necesitamos encontrar el valor de z correspondiente a esta media muestral, $\overline{X}$:

$$z = \frac{\overline{X} - \mu}{s_{\overline{x}}} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

$$z = \frac{51.5 - 50}{5/\sqrt{36}} = 1.8 \qquad (5)$$

Gráfico 36: Regiones de aceptación y rechazo en la prueba de hipótesis de sacos de 50 kilos con $H_0: \mu = 50$



Dado que este valor calculado de z, 1.8, cae dentro del área de aceptación, la compañía acepta H_0, es decir, acepta que el peso promedio de los sacos de azúcar que envía a sus clientes es de 50 kilos, con un nivel de confianza del 95%.

Una manera alternativa de establecer la regla de decisión es usar la unidad de medida de la variable; en este caso serán kilos. Así, queremos hallar la media muestral asociada al nivel de significación $\alpha = 5\%$, con dos colas, la que denotaremos como $\overline{X}_{\alpha/2}$ Anteriormente hemos determinado el *valor crítico* de $Z_{x/2} = 1.96$; entonces, despejando la ecuación (5) encontramos los valores críticos de $\overline{X}$:

$$\overline{X}_{\alpha/2} = \mu + Z_{\alpha/2}\, s_{\bar{x}} = 50 + 1.96\,(5/\sqrt{36}) = 51.63 \text{ kilos.}$$
$$\overline{X}_{\alpha/2} = \mu - Z_{\alpha/2}\, s_{\bar{x}} = 50 - 1.96\,(5/\sqrt{36}) = 48.37 \text{ kilos.}$$

Estos valores críticos de $\overline{X}_{\alpha/2}$ nos permiten establecer la *regla de decisión* en la unidad de medida de la variable bajo análisis; es decir, kilos:

Si $\overline{X} > 51.63$ ó $\overline{X} < 48.37$, rechazamos H_0.
Si $48.37 \leq \overline{X} \leq 51.63$, aceptamos H_0.

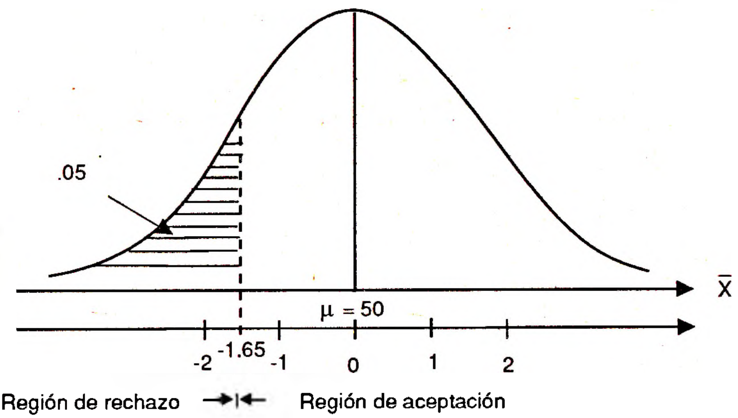
Dado que el peso promedio de los 36 sacos de la muestra es 51.5, valor que cae dentro del área de aceptación de la hipótesis nula, concluimos que la información muestral no evidencia que la media poblacional, el peso promedio de todos los sacos de azúcar que son enviados a los clientes, sea diferente a 50 kilos.

Ejemplo 2: Muestra grande, prueba de una cola

En el ejemplo de los sacos de azúcar de 50 kilos, supóngase ahora que el decisor es el Jefe de Almacén de una cadena de supermercados, y desea determinar con un nivel de confianza del 95% que los sacos de azúcar que compra no pesen menos de 50 kilos. Para esto toma una muestra aleatoria de n = 40 sacos y se halla que la media muestral es de 48 kilos y la desviación muestral, s, es igual a 5 kilos. Puesto que el Jefe de Almacén estará satisfecho si los sacos de azúcar pesan en promedio 50 kilos o más, establecerá las siguientes hipótesis:

(i) $H_0 : \mu \geq 50$ kilos
$H_1 : \mu < 50$ kilos
(ii) Nivel de significación $\alpha = 5\%$. Se usará la distribución z dado que $n>30$; y será una prueba de una cola, dado que la región de rechazo se encuentra en el extremo inferior de la distribución (véase el gráfico 37). El valor crítico de z será: $z_\alpha = -1.645$.
(iii) Regla de decisión
Si $z < -1.645$, rechazamos H_0.
Si $z \geq -1.645$, aceptamos H_0.

Gráfico 37: Regiones de aceptación y rechazo en la prueba de hipótesis de sacos de 50 kilos con $H_0: \mu \geq 50$



(iv) Dado que la media de la muestra de 40 sacos es 48 kilos, con una desviación estándar de 5 kilos, calculamos el valor de z correspondiente:

$$z = \frac{\bar{X} - \mu}{s_{\bar{x}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{48 - 50}{5/\sqrt{40}} = \frac{-2}{0.79} = -2.53$$

Puesto que este valor de z cae dentro de la región de rechazo ($z = -2.53 < -1.645$), debemos rechazar H_0, concluyendo, con un nivel de confianza del 95%, que el promedio de todos los sacos de azúcar recibidos es menor que 50 kilos. Esto llevará al Jefe de Almacén a rechazar el cargamento.

Al igual que en el ejemplo anterior, podemos establecer la regla de decisión fijando los valores críticos en kilos. Ahora queremos hallar $\bar{X}_\alpha$, asociada a $\alpha = 5\%$, pero con una sola cola. Dado $z_\alpha = -1.645$, entonces:

$$\overline{X}_\alpha = \mu - z_\alpha s_{\overline{x}} = 50 - 1.645\ (0.79) = 48.7 \text{ kilos}$$

La regla de decisión será, entonces:

Si $\overline{X} < 48.7$ kilos, rechazamos H_0 en favor de H_1.

Si $\overline{X} \geq 48.7$ kilos, aceptamos H_0.

Ejemplo 3: Muestra pequeña

Al igual que en el caso de muestras grandes, cuando se tiene una muestra pequeña, donde $n<30$, podemos realizar pruebas de hipótesis con una o dos colas, dependiendo del problema específico. La única diferencia es que en lugar de usar la distribución normal estandarizada se usará la distribución t de Student, siempre y cuando la población de la que se extrae la muestra tenga una distribución normal o aproximadamente normal.

Como ejemplo, suponga que una firma quiere saber, con un nivel de confianza del 99%, si puede garantizar que las baterías que producen duran más de 24 meses. Por experiencia, se sabe que la vida útil de las baterías tiene una distribución aproximadamente normal. La firma tomó una muestra aleatoria de 22 baterías y las entregó a chóferes de taxi para su uso. Encontró que la vida útil media de esta muestra de baterías fue de 25 meses con una desviación estándar de 3.8 meses. Puesto que es más conveniente incluir la igualdad en la hipótesis nula, y dado que el decisor desea probar que $\mu>24$ meses, debemos especificar nuestras hipótesis de la siguiente manera:

(i) $H_0 : \mu \leq 24$
$H_1 : \mu > 24$

(ii) Nivel de significación: $\alpha = 1\%$. Prueba de una cola; se usará la distribución t con $(n - 1) = 21$ grados de libertad, dado que la distribución de la población es aproximadamente normal, $n < 30$, y no se conoce σ.

(iii) Regla de decisión. El valor crítico de t con 21 grados de libertad y un nivel de significación de 1% es: $t(21.1\%) = 2.518$,

de la tabla en el apéndice C. Entonces, la regla de decisión está dada por:

Si $t > 2.518$, rechazamos H_0 en favor de H_1.

Si $t \leq 2.518$, aceptamos H_0.

(iv) El valor de t calculado para los estadígrafos de la muestra será:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{25 - 24}{3.8/\sqrt{22}} = \frac{1}{0.81} = 1.23$$

Como este valor de t es menor que el valor del t crítico, aceptamos H_0 con un nivel de confianza de 99%. Es decir, la compañía no puede afirmar que sus baterías duran más de 24 meses, a pesar de que la media muestral fue de 25 meses.

B. Prueba de hipótesis sobre la proporción de la población

La prueba de hipótesis sobre proporciones se usa cuando queremos determinar si la proporción de los elementos en una población que tiene cierta característica es mayor, igual o menor que algún valor específico. La lógica del procedimiento es idéntica a la establecida para la prueba de hipótesis de la media. Suponemos inicialmente que la hipótesis nula es cierta, determinamos la distribución muestral del estadígrafo (p) bajo el supuesto de que H_0 es cierta, escogemos un nivel de significación que determina una región crítica de rechazo de H_0, y finalmente tomamos la muestra. En lugar de utilizar los resultados de los teoremas 1 y 2, usaremos los teoremas 1' y 2' del capítulo anterior. Para el caso de la proporción, el valor de Z se calcula como:

$$z = \frac{p - \pi}{\sigma_p} \qquad (6)$$

donde π Proporción hipotética para el universo.
p Proporción en la muestra.
σ_p El error estándar de la proporción, dado por:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad (7.a)$$

o, si n>0.05 N

$$\sigma_p = \sqrt{\frac{\pi\,(1-\pi)}{n}}\;\sqrt{\frac{N-n}{N-1}} \qquad (7.b)$$

Ejemplo 4: Prueba de hipótesis sobre proporciones

Un vocero del gobierno que está disertando sobre la política cambiaria asegura que más del 50% de los exportadores están satisfechos con el tipo de cambio establecido en el último mes. Un profesor de la universidad, defensor de la promoción de exportaciones como medio para lograr el crecimiento económico del país, no cree en la afirmación gubernamental. Toma una muestra aleatoria de 50 personas que figuran en la lista de exportadores y halla que 27 de ellos están satisfechos con la política cambiaria del último mes. ¿Sustentan los datos muestrales la afirmación gubernamental al nivel de significación del 5%?

La hipótesis original del profesor universitario es que no más del 50% de los exportadores están satisfechos con el tipo de cambio. Entonces estableceremos las siguientes hipótesis:

(i) $H_0 : \pi \leq 0.50$
$H_1 : \pi > 0.50$

(ii) $\alpha = 0.05$. Usaremos la tabla z, ya que se trata de una muestra grande (n>30); y será una prueba de una sola cola. Luego, el valor crítico de z_α está dado por z_α=1.645.

(iii) Regla de decisión:
Si $z > z_\alpha = 1.645$, rechazamos H_0.
Si $z \leq z_\alpha = 1.645$, aceptamos H_0.

(iv) Calculamos el valor de z correspondiente al estadígrafo $p = 27/50 = 0.54$.

$$z = \frac{p - \pi}{\sigma_p} = \frac{0.54 - 0.50}{\sqrt{(0.5)\,(0.5)/50}} = \frac{0.04}{0.071} = 0.56$$

Puesto que este valor es menor que el valor crítico, es decir, se encuentra en la región de aceptación, aceptamos H_0. Esto significa que no hay base estadística suficiente para la afirmación gubernamental de que más del 50% de los exportadores se encuentran satisfechos con la política cambiaria del gobierno, al nivel de significación del 5%.

Debemos señalar que un error común en la prueba de proporciones es el uso del valor muestral p en lugar del valor hipotético de π en la determinación del error estándar de p. Recuerde que el error estándar p está dado por : $\sqrt{\pi\,(1 - \pi)/n}$, donde π es el valor hipotético en H_0. Si fijamos la regla de decisión antes de calcular el estadígrafo, p, evitaremos este error.

4. MEDIDA DEL ERROR DE TIPO II

Además del error de tipo I discutido en la sección anterior, existe un error de tipo II, que consiste en aceptar una hipótesis falsa. En un mundo de certidumbres siempre podríamos evitar cometer ambos tipos de errores, pero en el mundo real la evidencia muestral no es completamente confiable, y estamos sujetos a error. Una manera de enfrentar esta situación es decidiendo cuál es la probabilidad de error que estamos dispuestos a aceptar en la prueba de hipótesis. Naturalmente, nos gustaría que la probabilidad de incurrir en ambos tipos de error sea la más pequeña

posible. Pero para reducir ambos tipos de error se requiere seleccionar muestras más grandes, y es posible que estemos limitados tanto por recursos económicos como por disponibilidad de tiempo. Para un tamaño muestral específico, nos encontraremos en la posición de reducir la probabilidad de un tipo de error únicamente a costa de aumentar la probabilidad del otro error.

En la sección anterior nos concentramos en desarrollar un procedimiento para probar una hipótesis a un nivel de significación dado. Es decir, establecimos reglas de decisión en las cuales la probabilidad de cometer el error de tipo I –rechazar la hipótesis nula cuando era cierta– estaba determinada por un valor preestablecido de α. Como se señaló en la sección 1, una regla de decisión así establecida implicará necesariamente alguna probabilidad de cometer el error de tipo II, es decir, aceptar una hipótesis nula cuando esta es falsa. En esta sección consideramos la manera de calcular la probabilidad de cometer el error de tipo II, β.

Supongamos que tenemos una muestra aleatoria de n elementos de una población con media μ, desconocida, y desviación estándar σ, conocida, y queremos establecer las siguientes hipótesis:

$H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$

La regla de decisión para aceptar o rechazar H_0 depende del nivel de significación, del tamaño muestral y del tipo de prueba. Si suponemos que n>30, usaremos la distribución normal estandarizada. Puesto que se trata de una prueba de una sola cola, la regla de decisión será:

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\overline{X} - \mu_0}{\sigma \sqrt{n}} < z_\alpha$$

Por lo tanto, la probabilidad de rechazar la hipótesis nula cuando esta es cierta es igual al nivel de significación:

$$p\left(\frac{\overline{X}-\mu_0}{\sigma\sqrt{n}} < z_\alpha\right) = \alpha$$

Ahora estamos interesados en determinar la probabilidad de aceptar esta hipótesis nula cuando es falsa. Si la hipótesis nula es en efecto falsa, quiere decir que existe otra población alterna con una media poblacional realmente más pequeña que μ_0. A esta media alterna la denotaremos como μ_a. Supongamos que esta población alterna tenga una desviación estándar, σ_a; la variable aleatoria z_β tambien tendrá una variable normal estandarizada alrededor de μ_a, definida por:

$$z_\beta = \frac{\overline{X}-\mu_a}{\sigma_a/\sqrt{n}}$$

Este hecho nos permite calcular la probabilidad de cometer el error de tipo II, β, al aceptar H_0 cuando la población de la cual fue extraída la muestra es realmente una población alterna con media μ_a:

β = p (Aceptar la hipótesis nula cuando es falsa)

$$\beta = p\left(\frac{\overline{X}-\mu_a}{\sigma_a / \sqrt{n}} > z_\beta\right)$$

donde z_β mide la distancia que existe entre $\overline{X}$ y la media alterna, μ_a, en unidades de desviación estándar. Esta probabilidad puede calcularse usando las tablas de la distribución normal estandarizada, una vez que μ_a, σ_a, n y z_α son especificadas. Esto quiere decir que para calcular β es imprescindible especificar una población alterna con su respectiva media y desviación estándar.

Para ilustrar el cálculo de β, consideremos el ejemplo 2 de la sección anterior, en el que sometimos a prueba la hipótesis nula según la cual el peso promedio de los sacos de azúcar del cargamento recibido no era menor de 50 kilos:

H_0: $\mu \geq \mu_0 = 50$ kilos
H_1: $\mu < 50$ kilos

Dado que el tamaño de la muestra aleatoria es de 40 sacos, usamos la desviación estándar muestral, s, como estimado de σ. La prueba se llevó a cabo con un nivel de significación del 5% y se encontró el valor crítico de $z_\alpha = -1.645$. Se estableció la regla de decisión como:

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\overline{X}_\alpha - \mu_0}{\sigma/\sqrt{n}} < z_\alpha = -1.645$$

o, alternativamente, en términos de las unidades de medida de la media muestral, despejando $\overline{X}_\alpha$:

$$\text{Rechazar } H_0 \text{ si } \overline{X}_\alpha < 50 - 1.645\ (5/\sqrt{40}) = 48.7 \text{ kilos}$$

Es decir, el valor crítico de la media muestral es 48.7 kilos.

Ahora podemos determinar la probabilidad de que nuestra regla de decisión nos lleve a aceptar la hipótesis nula cuando el peso promedio verdadero de los sacos del cargamento sea en realidad μ_a, 47 y no 50 kilos.

$$\beta = p\ \left(\frac{\overline{X}_\alpha - \mu_a}{s/\sqrt{n}} > z_\beta \right)$$

$$\beta = p \left(\frac{48.7 - 47}{5/\sqrt{40}} > z_\beta \right) = p\ (2.15 > z_\beta)$$

De la tabla de la distribución normal estandarizada tendremos:

$$\beta = p\ (2.15 > z_\beta) = 0.016$$

Luego, concluimos que la probabilidad de aceptar la hipótesis nula, cuando el peso medio verdadero de los sacos del cargamento es 47 kilos, es de 1.6%.

El gráfico 38 ilustra estos cálculos. Además, podemos observar que si se reduce la probabilidad de cometer el error de tipo I (α), se aumenta el área de aceptación ($\overline{X}_\alpha$ se mueve a la izquierda) y, en consecuencia, aumenta la probabilidad de cometer el error de tipo II (β). Así, si fijamos α = 1%, entonces los valores críticos de z y $\overline{X}$ serán:

$$z_\alpha = -2.33$$

$$\overline{X}_\alpha = 50 - 2.33\ (0.79) = 48.1 \text{ kilos}$$

Entonces:

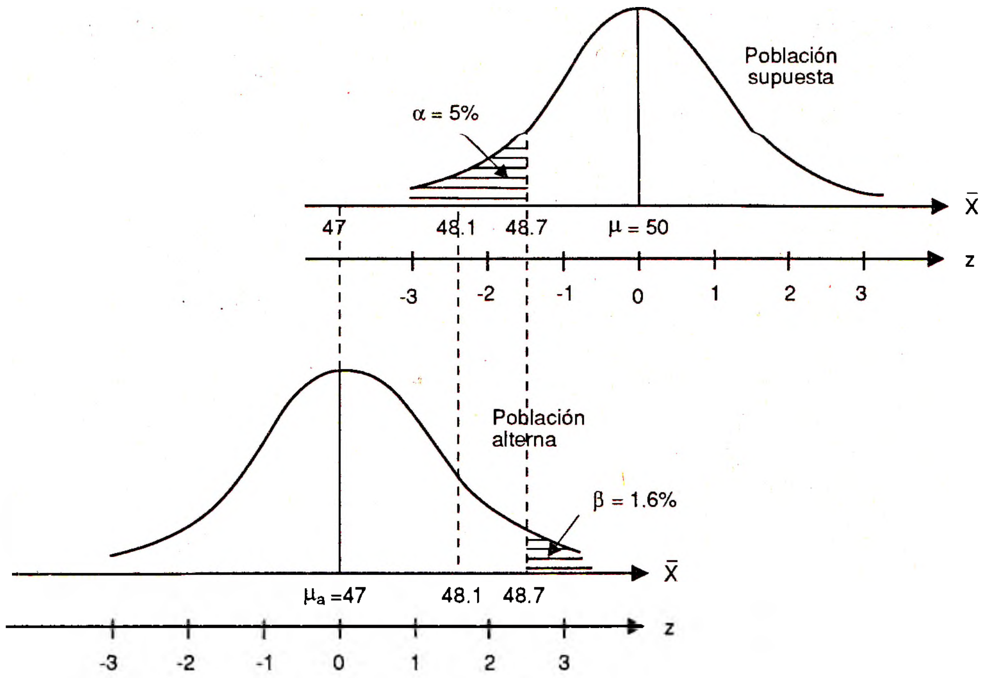
$$\beta = p\ (\overline{X} > 48.1) = p\ (z > \frac{48.1 - 47}{0.79}) = p\ (z > 1.39)$$

$$= 0.5 - 0.4177 = 0.08234$$

Es decir que si α = 1%, hay una probabilidad mayor de cometer el error de tipo II (β = 8.23%) que cuando α se fijó en 5%.

Si queremos mantener la probabilidad de cometer el error de tipo II en 1.6% y disminuir la probabilidad de cometer el error de tipo I de 5% a 1%, podemos establecer las regiones críticas para cada una de las dos curvas: la hipotética con μ_0 = 50 kilos y la alterna con μ_a = 47 kilos, definiendo $\overline{X}$ como el punto donde α=0.01 y β = 0.016.

Gráfico 38: Probabilidad de cometer el error de tipo II



De la curva de la población hipotética, con $\alpha = 0.01$, tenemos:

$$z_\alpha = -2.33 = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{\overline{X} - 50}{5/\sqrt{n}}$$

de donde:

$$\overline{X} = 50 - 2.33\,(5/\sqrt{n}) \qquad (8)$$

De manera similar, de la curva de la población alterna con $\beta = 0.016$, tenemos:

$$z_\beta = -2.13 = \frac{\overline{X} - \mu_a}{s/\sqrt{n}} = \frac{\overline{X} - 47}{5/\sqrt{n}}$$

de donde:

$$\overline{X} = 47 + 2.13\,(5/\sqrt{n}) \qquad (9)$$

Puesto que el valor de $\overline{X}$ es el mismo en las ecuaciones (8) y (9), tenemos:

$$50 - 2.33\,(5/\sqrt{n}) = 47 + 2.13\,(5/\sqrt{n})$$

Resolviendo para n, se obtiene un valor de $n = 55.25$.

Por lo tanto, si se incrementa el tamaño muestral de 36 a 55 sacos, será posible controlar las probabilidades de cometer los errores de tipo I y II en 1% y 1.6%, respectivamente. Esto ha sido posible porque al aumentar el tamaño muestral la variación en la distribución de las medias muestrales disminuye, dado que la desviación de las medias es $\sigma/\sqrt{n}$.

5. PRUEBA DE HIPÓTESIS PARA LA DIFERENCIA ENTRE DOS MEDIAS O DOS PROPORCIONES

En muchas situaciones de toma de decisiones, como en los controles de calidad, es importante determinar si las medias o

proporciones de dos poblaciones son o no iguales. Para hacer esto tomamos una muestra aleatoria de cada población, en forma independiente, y calculamos sus medias. Si los valores de las medias muestrales difieren, podríamos concluir que las muestras provienen de dos poblaciones con parámetros diferentes, o, alternativamente, que provienen de poblaciones que tienen parámetros iguales y que la diferencia entre los estadígrafos sólo se debe al azar. Luego, sólo si la diferencia entre las medias muestrales o proporciones se puede atribuir al azar aceptamos la hipótesis de que las dos poblaciones tienen iguales medias ($H_0 : \mu_1 = \mu_2$) o proporciones ($H_0 : \pi_1 = \pi_2$).

Estas pruebas son útiles para contestar preguntas como: ¿Hay diferencia en la duración de dos marcas diferentes de un producto dado? ¿Hay diferencia en las preferencias políticas entre los ciudadanos de Lima y Arequipa? ¿Son los salarios diferentes en dos empresas dadas?

A. Prueba para la diferencia entre dos medias

Para elaborar una prueba con respecto a la diferencia entre medias poblacionales, lo primero que hacemos es aplicar el resultado del teorema del límite central para este caso: Si tomamos muestras independientes de las dos poblaciones, de tamaño suficientemente grande (tanto n_1 como $n_2 \geq 30$), entonces la distribución muestral de la diferencia entre las medias muestrales $(\overline{X}_1 - \overline{X}_2)$ es normal o aproximadamente normal, definida por los siguientes parámetros:

$$E\ (\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 \tag{9}$$

y una desviación estándar o error estándar igual a:

$$\sigma\ (\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10}$$

donde μ_1 y σ_1 son la media y desviación estándar de la población de la cual se extrae el primer conjunto de muestras de tamaño n_1; μ_2 y σ_2 son la media y desviación estándar de la población de la cual se extrae el segundo conjunto de muestras de tamaño n_2. Podemos resumir estos resultados en la siguiente expresión:

$$(\overline{X}_1 - \overline{X}_2) \sim N(\mu_1 - \mu_2, \sigma_{(x_1 - x_2)})$$

Si no conocemos σ_1 y σ_2, y los tamaños muestrales son mayores que 30, usaremos las desviaciones muestrales (s_1 y s_2) como una buena aproximación de las desviaciones estándar poblacionales, y el error estándar de las diferencias será igual a:

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (11)$$

En el caso en que tanto n_1 como n_2 sean menores de 30; que σ_1 y σ_2 sean desconocidas; y que podamos suponer que las muestras provienen de poblaciones que tienen distribución normal, entonces la distribución muestral de la diferencia entre las medias tendrá una distribución t con $(n_1 + n_2 - 2)$ grados de libertad. Si suponemos que las desviaciones estándar poblacionales son iguales pero desconocidas, usamos s_1 y s_2 para estimar el error estándar de la diferencia entre medias de la siguiente manera:

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (12)$$

Donde:

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \qquad (13)$$

es decir, s^2 es un promedio ponderado de s_1^2 y s_2^2. Las ponderaciones son $(n_1 - 1)$ y $(n_2 - 1)$, con el fin de obtener estimaciones no-sesgadas para σ_1^2 y σ_2^2.

Ahora ya podemos aplicar el procedimiento de contraste de hipótesis presentado en la sección 3.

Ejemplo 5

Una importante compañía de transporte público de Lima Metropolitana debe decidir entre dos marcas de llantas para su parque automotor, con un nivel de confianza del 95%. Para tomar la decisión seleccionó una muestra aleatoria de 100 llantas de cada marca y encontró que la marca 1 tiene una vida útil de 98,000 kilómetros en promedio, $\overline{X}_1$, con una desviación estándar, s_1, de 8,000 kilómetros. Por otro lado, los estadígrafos calculados para la marca 2 son $\overline{X}_2$ = 101,000 kilómetros y s_2 = 12,000 kilómetros. ¿Qué marca de llantas debería adquirir la compañía de transporte si la diferencia de precios es mínima?

Aplicando el procedimiento de prueba de hipótesis, tenemos:

(i) $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

o, alternativamente:

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 - \mu_2 \neq 0$

(ii) α=5%. Usaremos la tabla z, pues tanto n_1 como n_2 son mayores de 30; es una prueba de dos colas, con la región de aceptación en el intervalo limitado por $Z_{\alpha/2} = \pm 1.96$.

(iii) Regla de decisión:
Si $z > 1.96$ ó $z < -1.96$, rechazamos H_0.

(iv) Usando los datos muestrales podemos calcular el valor de z, para lo cual primero calculamos la desviación estándar de la diferencia entre las medias muestrales:

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(8{,}000)^2}{100} + \frac{(12{,}000)^2}{100}} = 1{,}442 \text{ km}$$

Luego:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{x}_1 - x_2)}} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_{(x_1 - \bar{x}_2)}}$$

$$= \frac{98{,}000 - 101{,}000}{1{,}442} = \frac{-3{,}000}{1{,}442} = -2.08$$

Este valor calculado de z cae dentro de la región de rechazo de H_0 ($z = -2.08 < Z_{\alpha/2} = -1.96$), y la compañía de transporte debe aceptar $H_1 : \mu_1 \neq \mu_2$, al nivel de significación del 5%.

Hemos determinado que existe una diferencia significativa entre la vida útil media de ambas marcas. Sin embargo, no hemos contestado nuestra pregunta inicial de qué marca de llantas se debe adquirir. Realizamos una nueva prueba de hipótesis suponiendo que la vida útil media de las llantas de marca 2 es mayor que la de la marca 1.

H_0: $\mu_1 \geq \mu_2$
H_1: $\mu_1 < \mu_2$

Usando el mismo α, la tabla z, y dado que es una prueba de una sola cola, el valor crítico de z_α estará dado por - 1.645.

De los datos muestrales calculamos que z = -2.08, que es menor que el z crítico, por lo que rechazamos H_0 y concluimos que la vida útil promedio de las llantas de la marca 2 es mayor que aquella de la marca 1. Por lo tanto, la compañía de transporte debe abastecerse de la marca 2.

Para resaltar la importancia de la elección del nivel de significación en el proceso de prueba de hipótesis, supongamos que

el decisor de la compañía opta por un nivel de significación del 1% en lugar del 5%. Si:

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

el valor crítico de z será igual a $z_{\alpha/2} = \pm 2.58$, y la regla de decisión será:

Rechazar H_0 si $z > 2.58$ ó $z < -2.58$.

El z calculado ($z = -2.08$), al nivel de significación del 1%, caerá dentro de la región de aceptación de H_0. Esto indicaría que no hay diferencia significativa entre μ_1 y μ_2 al nivel de significación del 1%, así que la compañía de transporte podría adquirir cualquiera de las dos marcas.

B. Prueba para la diferencia entre dos proporciones

La prueba de hipótesis para la diferencia entre dos proporciones se realiza cuando queremos determinar si las proporciones de dos poblaciones son o no iguales. La lógica del procedimiento es idéntica a la establecida para la diferencia de las medias. Tomamos una muestra aleatoria de cada población y calculamos las proporciones muestrales; si la diferencia entre estas proporciones se puede atribuir al azar, aceptamos la hipótesis de que las dos poblaciones tienen iguales proporciones.

El teorema del límite central para el caso de la diferencia entre proporciones establece que si tomamos muestras suficientemente grandes ($n_1, n_2 \geq 30$), entonces la distribución muestral de la diferencia entre las proporciones es normal:

$$(p_1 - p_2) \sim N(\pi_1 - \pi_2, s_{(p_1 - p_2)}) \quad (14)$$

donde:

$$s_{(p_1 - p_2)} = \sqrt{p\,(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{15}$$

Para calcular el error estándar de la diferencia entre proporciones, $s_{(p_1 - p_2)}$, necesitamos un estimado de la proporción total de éxitos para las dos muestras combinadas, al que llamaremos p:

$$p = \frac{n_1\,p_1 + n_2\,p_2}{n_1 + n_2} \tag{16}$$

donde p_1 es la proporción de éxitos en la muestra 1, y p_2 es la proporción en la muestra 2. Es decir, p es un promedio ponderado de p_1 y p_2.

Ejemplo 6

Una compañía quiere determinar con un nivel de significación del 5% si la proporción de empleados del Departamento de Producción que llega tarde, π_1, es mayor que dicha proporción en el Departamento de Administración, π_2. La compañía toma una muestra aleatoria de cada uno de estos departamentos y encuentra que $p_1 = 0.08$ y $p_2 = 0.05$ para $n_1 = 40$ y $n_2 = 30$.

Puesto que la compañía quiere probar si $\pi_1 > \pi_2$, plantearemos las siguientes hipótesis:

(i) $H_0 : \pi_1 \leq \pi_2$
$H_1 : \pi_1 > \pi_2$

o, alternativamente:

$H_0 : \pi_1 - \pi_2 \leq 0$
$H_1 : \pi_1 - \pi_2 > 0$

(ii) $\alpha = 5\%$. Usamos la tabla z, ya que $n_1, n_2 > 30$; es una prueba de una cola, con la región de rechazo en el límite superior de la distribución. Luego, el z crítico será $z_\alpha = +1.645$.

(iii) Regla de decisión:

Si $z > 1.645$, rechazamos H_0.

(iv) De los datos muestrales, tenemos:

$$p = \frac{400\,(0.08) + 200\,(0.05)}{400 + 200} = \frac{10}{600} = 0.017$$

$$s_{(p_1 - p_2)} = \sqrt{(0.017)\,(0.983)\left(\frac{1}{400} + \frac{1}{200}\right)} = 0.0112$$

Entonces:

$$z = \frac{p_1 - p_2}{s_{(p_1 - p_2)}} = \frac{0.08 - 0.05}{0.0112} = 2.68$$

Puesto que $z = 2.68 > z_\alpha = 1.645$, rechazamos H_0 y concluimos, al nivel de significación del 5%, que la proporción de tardanzas en el Departamento de Producción es mayor que la del Departamento de Administración.

● *Ejercicios*

1. Defina y discuta los siguientes conceptos:

a. Error de tipo I y error de tipo II.

b. Hipótesis nula e hipótesis alternativa.

c. Pruebas de una y dos colas.

d. Regla de decisión del procedimiento de prueba de hipótesis.

e. Parámetros α y β.

2. Un productor nacional de refrigeradoras sostiene que su producto tiene en promedio una vida de 30 meses sin desperfectos. En una muestra de 85 refrigeradoras, la vida promedio antes de necesitar una

reparación fue de 28.3 meses, con una desviación estándar de 5 meses. ¿Proporciona esta muestra suficiente evidencia para refutar la afirmación del fabricante? Use un nivel de significación de 0.05.

3. La vida promedio de una muestra de 100 tubos fluorescentes producidos por una compañía fue de 1,570 horas con una desviación estándar de 120 horas. El Gerente de la compañía afirma que sus tubos fluorescentes duran por lo menos 1,600 horas. Usando el nivel de significación de 0.05, ¿está usted de acuerdo con la afirmación del Gerente?

4. Un fabricante está considerando la compra de un nuevo equipo para elaborar herramientas y ha especificado que el equipo no debe requerir, en promedio, más de 10 minutos de preparación por cada hora de operación. El encargado del Departamento de Compras visita una compañía donde se halla instalado el equipo y comprueba que 40 horas de operación seleccionadas al azar incluyen 7 horas y 30 minutos de preparación con una desviación estándar de 3 minutos. Basándose en estos resultados para la muestra, ¿puede rechazarse la suposición de que el equipo posee las especificaciones sobre tiempo de preparación al nivel de significación del 1%?

5. Se ha creado un nuevo tipo de cuenta "Iffy" para aquellas personas que tienen un historial de crédito malo y que no son elegibles para otro tipo de cuenta en el banco. Un cliente que posee una cuenta "Iffy" debe pagar una multa de 5% mensual sobre los montos atrasados por más de un mes. Si no se cancelan los créditos pendientes en tres meses, se suspende la cuenta. El encargado de las cuentas "Iffy" sostiene que el balance promedio no pagado es usualmente igual o menor a $ 200. Una muestra de veinte cuentas impagas revela que el balance promedio es de $ 210 con una desviación estándar de $ 25.

a. ¿Es justificada la afirmación del encargado de las cuentas "Iffy"? Use un nivel de significación de 0.05.

b. Si se tomara una muestra de 200 cuentas y se obtuviera una media de $ 210 y una desviación estándar de $ 25, ¿cuál sería su conclusión?

6. Se piensa que los valores que se muestran a continuación proceden de una población con media igual a 20. Pruebe esta hipótesis utilizando un nivel de significación del 0.05%. Los valores son: 16, 14, 24, 18, 22 y 20.

7. Los accionistas de una cadena de tiendas por departamentos están dispuestos a establecer una nueva sucursal en un cierto distrito si el ingreso familiar promedio en esa área es de por lo menos $ 1,000 mensuales. Se realiza una encuesta en 400 hogares seleccionados al azar, encontrando que el ingreso familiar mensual promedio es $ 986, con

una desviación estándar de $ 124. Utilice un nivel de significación de 0.01 para determinar si los datos de la muestra justifican que se establezca una nueva sucursal en ese lugar.

8. Considere al productor de focos "Brillante". La campaña publicitaria afirma que el producto tiene una "vida de 100 horas o más". La vida de las bombillas tiene una distribución normal con una desviación estándar de 5 horas. Un lote de producción consiste de 7,000 a 8,000 focos.

a. Construya una regla de decisión para una muestra aleatoria de 25 focos, al nivel de significación del 10%.

b. Calcule la probabilidad de aceptar la hipótesis nula utilizando la regla de decisión de la parte (a) cuando la "media verdadera de la población" es de 99 horas con la misma desviación estándar especificada en (a).

9. Cierto fabricante de refrescos alega que su producto es superior al de su competidor porque en una muestra de 100 personas que probaron ambos refrescos con los ojos vendados, el 52% prefirió su producto. ¿Justifica este resultado muestral la afirmación del fabricante? Utilice un nivel de significación de 0.05.

10. Durante veinte días se registraron las temperaturas de operación de dos hornos de secado para pintura. Las medias y las varianzas de las dos muestras son:

	Media	Varianza
Horno A	82	41
Horno B	84	86

¿Proporcionan estos datos evidencia suficiente para indicar que existe una diferencia en la temperatura promedio de secado para los dos hornos? Use un nivel de significación de 0.05.

11. El laboratorio farmacéutico "El Inca", que fabrica "Inmejorable", unas pastillas para eliminar el dolor de cabeza, afirma que su producto produce alivio más rápido que el de la competencia. Para fundamentar su argumento tomó dos muestras de personas con dolor de cabeza a quienes se les administró los analgésicos en cuestión. Se obtuvieron los siguientes resultados:

	"Inmejorable"	Producto de la competencia
Tiempo promedio que demora el producto en comenzar a tener efecto	12 minutos	14 minutos
Desviación estándar de la muestra	4 minutos	3 minutos
Tamaño de la muestra	100	100

¿Es esta evidencia consistente con la afirmación del laboratorio? Utilice un nivel de significación de 0.05.

12. Una empresa que fabrica e instala calentadores solares para viviendas está considerando iniciar operaciones en Chaclacayo. La compañía estima que le será conveniente iniciar operaciones en ese lugar si el 50% ó más del total de viviendas están habitadas por sus dueños. Una encuesta de 31 hogares de Chaclacayo indica que el 58% de las viviendas son alquiladas. Con un nivel de significación de 0.05, ¿debería la empresa iniciar operaciones en Chaclacayo?

13. Una compañía de publicidad sostiene que el 40% de los automovilistas que cruzan cierta intersección ven el anuncio allí ubicado y recuerdan su mensaje. Se toma una muestra al azar, formada por 640 automovilistas que pasaron por dicho lugar, de los cuales 224 recordaron el anuncio. ¿Es posible rechazar la afirmación de la compañía de publicidad? Utilice un nivel de significación de 0.01 con una prueba de una cola.

14. Se lanza un dado 240 veces y se obtiene el número "6" cincuenta veces. ¿Se justifica la afirmación de que el dado tiende a favorecer al seis, con un nivel de significación del 5%?

15. La experiencia de la tienda "Pretty" indica que el 70% de sus clientes son mujeres. "Pretty" inicia una nueva política de mantenerse abierta los domingos y toma una muestra aleatoria de 896 de sus clientes dominicales, encontrando que el 66% de ellos son mujeres. Determine, a un nivel de significación de 0.05, si este hallazgo indica que la proporción de mujeres entre los clientes dominicales difiere de la proporción de los otros días.

16. Un vendedor de "Electrolux" encontró que sus ventas por cliente eran de $ 350 con una desviación estándar de $ 80. Después de un aumento significativo en avisos promocionales en el mes de octubre, tomó una muestra aleatoria de 40 facturas y se encontró que

el valor de venta promedio por cliente fue de $ 370 con la misma desviación estándar de $ 80. Determine la efectividad de la campaña de publicidad sobre el promedio de ventas con un nivel de significación de 0.05.

17. El Jefe de la Oficina de Personal de una gran cervecería realizó un estudio para calcular los años de servicio de los 10,000 empleados de la compañía como parte de una investigación para determinar si era deseable establecer un nuevo plan de retiro. El estudio encontró que el promedio (media aritmética) de los años de servicio era de 12.5 años, con una desviación estándar de 4 años. Cinco años más tarde se tomó una muestra aleatoria de 324 empleados y se encontró que el promedio de años de servicio en la muestra era de 13.8 años.

Determine, con un nivel de significación de 0.01, si la información en la muestra indica que ha aumentado el número promedio de años de servicio de los empleados de la cervecería. Suponga que la desviación estándar permanece constante.

18. Cierta máquina que llena sacos de alimentos balanceados para animales menores está calibrada para embolsar 45 kilos de alimento en cada saco. Se tomó una muestra aleatoria de cinco sacos y se pesó su contenido en kilos: 44.2, 46.3, 45.1, 47.2 y 45.8 respectivamente. Utilice un nivel de significación de 0.05 para analizar si la máquina está bien calibrada.

19. Los 84 estudiantes que solicitaron admisión en la Escuela de Postgrado de la Universidad del Pacífico en este ciclo obtuvieron un puntaje promedio de 67 en el examen de admisión, con una desviación estándar de 21. En el ciclo anterior, los 68 estudiantes que solicitaron admisión obtuvieron una calificación promedio de 70, con una desviación estándar de 35.

a. ¿Se puede afirmar que los postulantes de este ciclo tienen un rendimiento inferior a los del ciclo anterior a un nivel de significación de 0.01?

b. ¿Cuál es la región de aceptación para la prueba en términos de los resultados del examen de admisión?

20. La compañía manufacturera de amortiguadores asegura a sus distribuidores que no encontrarán más del 6% de amortiguadores defectuosos en cada envío. Un distribuidor decide poner a prueba esta afirmación usando un 0.05 de probabilidad de rechazar la afirmación si esta es cierta, dado que el productor sólo aceptaría una tasa de devolución del 5% de mercadería "buena". El distribuidor escogió 125 amortiguadores aleatoriamente, y encontró 11 defectuosos. ¿A qué conclusión se puede llegar respecto a la afirmación del productor?

21. "Los Perforadores", contratistas que realizan trabajos de perforación de pozos petroleros, tienen una especial preocupación por utilizar sellos de calidad. Para aumentar el margen de seguridad, estos sellos se utilizan en grupos sucesivos de tres, pues el 20% de ellos fallan antes de su tiempo de reemplazo de 100 horas.

La firma adquirió recientemente un lote de 5,000 sellos de un nuevo fabricante y probó 225. Cincuenta y dos fallaron antes de su tiempo de reemplazo. ¿Pueden "Los Perforadores" concluir que los nuevos sellos son inferiores a los que habían estado usando? Justifique su respuesta.

22. Una compañía local que importa llantas ha recibido una serie de quejas acerca de la duración de las mismas. La empresa decidió tomar una muestra de 64 llantas del último embarque y someterlas a prueba. Dado que la prueba es destructiva, la empresa no quiso tomar una muestra muy grande. De acuerdo con los fabricantes, la duración promedio de las llantas debería ser por lo menos 40,000 kilómetros, con una desviación estándar conocida de 400.

La compañía encontró que la duración promedio de las llantas de la muestra es de 38,000 kilómetros. ¿Debería devolver el cargamento o aceptarlo? Use un nivel de significación de 0.05, dado que los abastecedores no aceptarían más del 5% de devoluciones de buena mercadería.

23. Una compañía manufacturera que produce dos tipos de llantas, una que contiene fajas de acero y la otra que usa partículas de plutonio, decidió realizar una prueba. Se distribuyeron llantas a una muestra de chóferes interprovinciales, midiendo el desgaste cada cierto tiempo. Confiando en los comentarios de los usuarios y en la medición del desgaste de las llantas, se elaboró la siguiente tabla:

Tipo de llanta	Vida media (km)	Desviación estándar	Número de usuarios
Acero	43,000	3,500	11
Plutonio	40,800	2,100	12

Con un nivel de significación del 5%, ¿concluiría que hay una diferencia entre los dos tipos de llantas? Si la hubiere, ¿qué tipo de llanta se debe producir, si es más barato fabricar las llantas de acero?

24. Cierta empresa está considerando introducir un nuevo plan de retiro si el promedio de los años de servicio de sus empleados es de alrededor de 12 años, pero estima que el nuevo plan sería demasiado costoso si el promedio de años de servicio excede los 14 años. De un estudio previo se determinó que la desviación estándar de los años de servicio es de 4 años. Diseñe un experimento que permita a la compañía tomar una decisión con $\alpha = 0.01$ y $\beta = 0.05$.

VI. Análisis de regresión y correlación

1. Diagramas de dispersión. 2. Ecuación de regresión-Método de mínimos cuadrados ordinarios. 3. Medida del error estándar de estimación. 4. Uso del error estándar de estimación. 5. Coeficiente de determinación y coeficiente de correlación. 6. Prueba de significación del coeficiente de determinación. 7. Prueba de significación del coeficiente de regresión. 8. Condiciones para el uso del método de mínimos cuadrados ordinarios. 9. Análisis de correlación lineal. 10. Análisis de regresión múltiple. 11. El análisis de regresión y la computadora. 12. Los supuestos del método de mínimos cuadrados ordinarios y métodos alternativos de estimación. 13. Limitaciones del análisis de regresión.

Este capítulo se dedicará al estudio del análisis de asociación o relación que pudiera existir entre dos o más variables. Podremos examinar, por ejemplo, la relación entre los niveles de importaciones de insumos y la producción industrial; entre las notas que obtienen los alumnos en el curso de Análisis Estadístico y en el curso de Matemáticas; entre la temperatura promedio, la precipitación de lluvia y la producción de papas en una región determinada.

Los dos métodos que se discutirán son el análisis de regresión y el análisis de correlación. Estos métodos difieren en su enfoque fundamental. El análisis de regresión considera una o más variables como dadas, y examina el efecto de estas en una variable aleatoria dependiente. El análisis de correlación considera a todas las variables como aleatorias, y examina la interrelación que existe entre ellas.

El análisis de regresión tiene como objetivo predecir el valor de una variable tomando como base a otras que la explican, medir el grado de relación entre estas variables y probar hipó-

tesis para establecer si la relación es o no significativa. Para lograr estos propósitos se utiliza el diagrama de dispersión que se explica en la sección 1, y la estimación de una ecuación lineal por el método de minímos cuadrados que se discute en la sección 2. En la sección 3 se establece la utilidad de esta ecuación para estimar valores de la variable independiente y medir la confiabilidad de tales estimados a tráves del cálculo del error estándar de estimación. Con base en este error estándar de estimación, que mide el grado de relación entre las variables, se construyen intervalos de confianza, como se muestra en la sección 4. Los coeficientes de determinación y correlación se calculan en la sección 5. Para terminar con el análisis de regresión simple, en las secciones 6 y 7 se introduce el uso de la prueba t para determinar si la relación entre las variables y el coeficiente de regresión es estadísticamente significativa. En la sección 8 se explicitan las condiciones que deben cumplirse para poder usar el método de mínimos cuadrados ordinarios en la estimación de una ecuación lineal.

En la sección 9 se discuten las medidas de interdependencia de dos variables definidas por el análisis de correlación. Luego, en las secciones 10 y 11 se amplía y generaliza el análisis de regresión, para considerar los efectos de más de una variable independiente en una variable dependiente; así como ciertas relaciones no lineales (sección 12). Finalmente, en la sección 13 se establecen algunas limitaciones del análisis de regresión.

1. DIAGRAMAS DE DISPERSIÓN

Consideremos el siguiente ejemplo: La compañía minera "La Hermosa" desea establecer la relación que existe entre su nivel de producción de minerales en toneladas métricas y el número de horas/hombre empleadas en dicha producción. Se plantea la hipótesis de que a medida que aumentan las horas empleadas, aumenta la producción. Antes de emplear las técnicas matemáticas del análisis de regresión para aceptar o rechazar esta hipó-

tesis, trataremos de representar los datos disponibles para las dos variables en un diagrama para verificar si existe tal relación. El Departamento de Producción de la mina proporcionó información sobre las variables de interés para el período enero-setiembre del año pasado. Estos datos se muestran en el cuadro 12.

CUADRO 12: COMPAÑÍA MINERA "LA HERMOSA": NIVELES DE PRODUCCIÓN DE MINERAL Y NÚMERO DE HORAS/HOMBRE EMPLEADAS

Mes	Miles de horas/hombre	Mineral producido (Miles de toneladas)
Enero	1,170	35.4
Febrero	1,150	33.1
Marzo	1,343	48.0
Abril	757	18.5
Mayo	1,180	36.4
Junio	1,117	36.0
Julio	1,180	34.1
Agosto	933	28.1
Setiembre	1,195	39.7

De la observación de estos datos podemos afirmar que, en general, el nivel de producción aumenta a medida que el número de horas/hombre empleadas para su producción se incrementa. La relación podría ser descrita como lineal. Cada mil horas/hombre adicionales tenderán a aumentar la producción de minerales en una cantidad constante. Sin embargo, esta relación lineal describe solamente una tendencia general; no dice exactamente cuánto se producirá con base en el número de horas/hombre empleadas.

La representación de los datos en un plano cartesiano se llama *diagrama de dispersión*. Por convención, la variable conocida o independiente se grafica en el eje de las abscisas (X), y la variable dependiente, a ser estimada, en el eje de las ordenadas (Y). Por lo tanto, cada punto de esta gráfica representa dos características

de la misma unidad de análisis. Esta representación gráfica de los datos es *informativa,* pues nos permite tener una idea sobre el grado (intensidad) y la naturaleza (forma) de la relación entre las dos variables. Por inspección podemos establecer si la relación es lineal o no-lineal, positiva o negativa, o si simplemente no existe una relación aparente. Para obtener la máxima información posible de un diagrama de dispersión se deberá tener especial cuidado en elegir tanto las unidades de medida de las variables en estudio, como las escalas que se utilizarán para graficarlas.

El diagrama de dispersión para el caso de la mina "La Hermosa" se muestra en el gráfico 39. Observando el diagrama, podemos establecer lo siguiente:

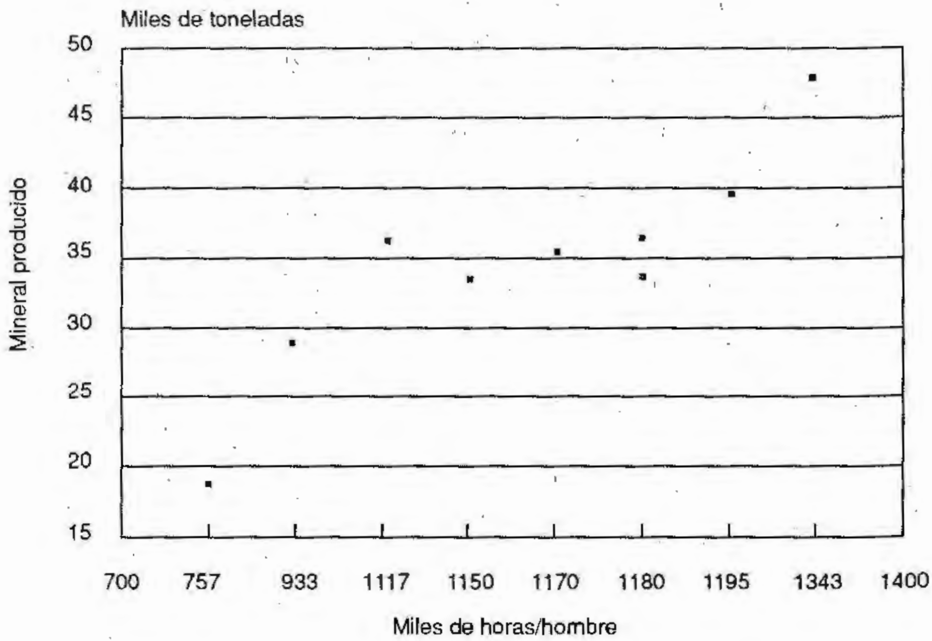
1. Existe una relación lineal entre el nivel de mineral producido en un mes y el número de horas/hombre empleadas en dicho período. Es posible trazar una línea recta que se "ajuste" a los puntos graficados en el diagrama de dispersión.

2. Es obvio que la relación no es determinística; vale decir, cualquiera sea la línea recta que se trace, la mayoría de los puntos estarán por encima o por debajo de dicha línea.

El punto 2 expresa que la relación entre las dos variables no quedará totalmente explicada por una ecuación lineal, sino que en cada mes habrá factores peculiares que harán que se produzca un poco más o menos de lo esperado si sólo se considera el número de horas/hombre empleadas.

El diagrama de dispersión no sólo es útil en el análisis de regresión, sino también en el análisis de correlación, para indicar el grado de asociación entre dos variables, como veremos en la sección 9. En este último caso, no se trata de una variable dependiente y otra independiente, pues las dos son consideradas aleatorias. En el análisis de regresión se supone que la variable explicativa o independiente toma valores fijos en muestras repetidas. Es decir, que los valores de X han sido medidos sin error, mientras que los valores de Y asociados a los diferentes valores de X sí son aleatorios.

Gráfico 39: Diagrama de dispersión entre el número de horas/hombre empleadas y el nivel de producción de mineral



2. ECUACIÓN DE REGRESIÓN-MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS (MCO)

En el caso de la mina "La Hermosa", se desea saber cuánto se producirá en un mes si se cuenta con un determinado número de horas/hombre. Como observamos en el diagrama de dispersión, existe una relación lineal entre ambas variables, y nuestro problema es ahora trazar una línea recta que se "ajuste" lo mejor posible a los puntos de dicho diagrama. Esta *línea de regresión*, que representa la relación promedio entre las dos variables, será el instrumento que nos permitirá estimar el valor de la variable dependiente (niveles de producción) tomando como base el valor de la variable independiente (horas/hombre).

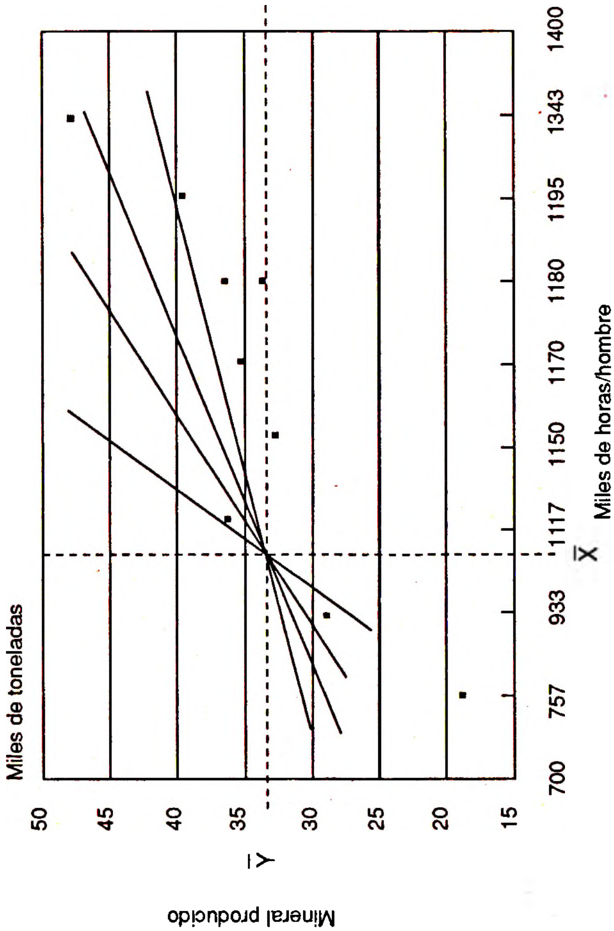
Necesitamos establecer un método para *estimar* o ajustar una línea de los datos observados. Además, debemos seleccionar criterios para medir la *bondad de ajuste*, estableciendo una medida de lo bien o mal que la línea de regresión logra describir la relación entre las variables.

Podríamos ajustar una línea, "al ojo", trazando una recta que consideremos sea la que mejor se ajusta. Otra persona podría trazar otra línea afirmando que se ajusta tan bien o mejor que la nuestra. No hay manera de decidir cuál es la línea que mejor se ajusta hasta que estemos de acuerdo en un criterio que defina lo que constituye un buen ajuste.

Otro método simplista podría especificar que la línea de ajuste pase por el punto $(\overline{X}, \overline{Y})$, siendo $\overline{X}$ e $\overline{Y}$ los valores de las medias de las variables. Pero como se puede observar en el gráfico 40, existen una infinidad de líneas que pasan por $(\overline{X}, \overline{Y})$.

A través del análisis de regresión buscamos que la línea de ajuste se aproxime lo más posible a todos los puntos del diagrama de dispersión. Algunos puntos estarán por encima de la línea, mientras que otros estarán por debajo. Sin embargo, esta línea debe ser *no-sesgada*, en el sentido de que no tienda a sobreestimar los puntos más que a subestimarlos. Como primer

Gráfico 40: Líneas que pasan por el punto ($\overline{X}$,$\overline{Y}$)



criterio de ajuste se podría especificar que la suma de las desviaciones de cada punto a la línea, llamada error total, sea mínimo.

La línea de regresión tendrá la forma:

$$\hat{Y}_i = a + bX_i$$

donde $\hat{Y}_i$ es el valor estimado de Y_i para X_i.
a es el intercepto en el eje Y.
b es la pendiente de la línea, conocida como el coeficiente de regresión.

El gráfico 41 muestra el ajuste de la línea Y a los puntos de nuestro ejemplo. La desviación o error es medida verticalmente del punto a la recta; es decir, la diferencia entre el valor actual, Y, y el valor estimado basándose en X, $\hat{Y}$. Luego, el error total será la suma de todos los errores:

$$\sum (Y_i - \hat{Y}_i) \qquad (1)$$

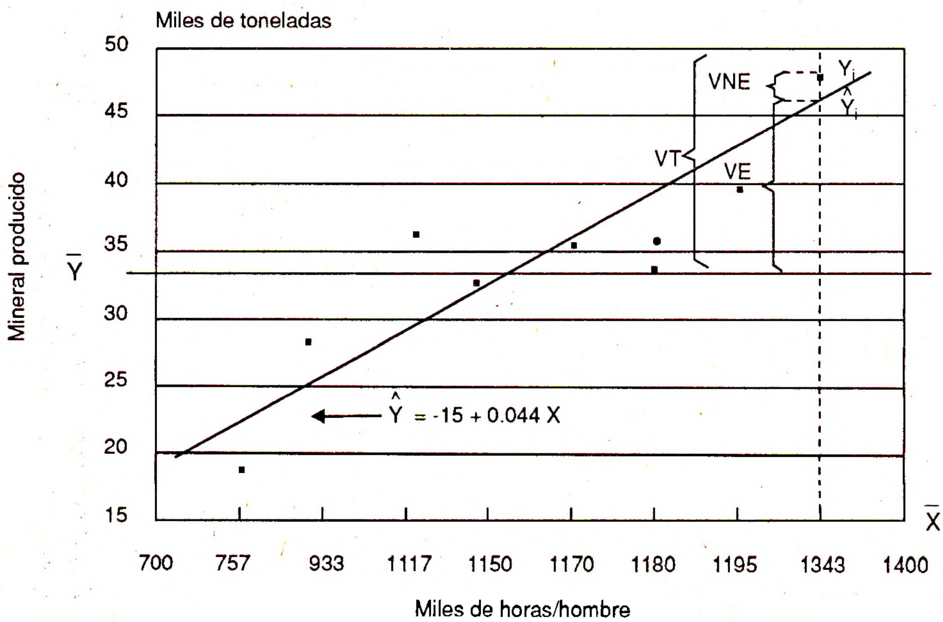
donde Y_i es el valor observado Y, asociado con X_i.
$\hat{Y}_i$ es el valor estimado para X_i usando la línea de regresión

Si especificamos que la línea de mejor ajuste es la que minimiza la suma de estos errores, y si esta línea pasa por el punto determinado por el valor de las medias de X e Y, $(\overline{X}, \overline{Y})$, los errores negativos y positivos se cancelarán por definición de la media. Además, como vimos anteriormente, podremos trazar una infinidad de líneas que pasan por $(\overline{X}, \overline{Y})$, y todas ellas tendrán un error total igual a cero.

Podemos eliminar el problema de las cancelaciones de las desviaciones positivas y negativas usando la suma de los valores absolutos de las desviaciones. Es decir, minimizando:

$$\sum | Y_i - \hat{Y}_i | \qquad (2)$$

Gráfico 41: Variación explicada (VE), variación no explicada (VNE) y variación total (VT)



Un mejor criterio para determinar el *mejor ajuste* para una relación lineal es el que minimiza la suma de errores al cuadrado:

$$\sum (Y_i - \hat{Y}_i)^2 \qquad (3)$$

Este criterio es más conveniente, puesto que además de eliminar el problema de la compensación de los errores, permite la manipulación matemática necesaria para estimar la línea de regresión. El método que se deriva de este criterio se conoce como el *método de mínimos cuadrados* (MC).

Como se indicó anteriormente, la línea de regresión que permite predecir el valor de Y, $\hat{Y}$, con base en los valores de X, se expresa como:

$$\hat{Y}_i = a + bX_i \qquad (4)$$

En el anexo que aparece al final de este capítulo se presenta la derivación matemática para determinar los valores de a y b que minimizan la suma de los errores al cuadrado. Estos valores están dados por las siguientes ecuaciones:

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \qquad (5)$$

$$a = \overline{Y} - b\overline{X} \qquad (6)$$

Aplicando estas fórmulas al caso de la mina "La Hermosa", obtendremos la siguiente línea de regresión:

$$\hat{Y} = -15.000 + 0.044X \qquad (7)$$

Interpretación de la ecuación lineal de regresión

Podemos graficar la línea de regresión estimada en el diagrama de dispersión original. Cuando X vale cero, la ecuación nos permite obtener el valor asociado $\hat{Y}$, ($\hat{Y}$ = -15.000); cuando X toma el valor mayor del rango de los datos, 1,343, el valor estimado de $\hat{Y}$ es 44.09. Si graficamos estos dos puntos en el diagrama de dispersión y los conectamos con una línea recta, obtenemos la representación gráfica de la ecuación de regresión estimada basándose en los nueve puntos observados.

Los valores de a y b son estadígrafos, puesto que se los calcula tomando como base la información muestral. El valor de *a* es el *intercepto en Y*, es decir, el valor de Y cuando X es cero. En términos de nuestro ejemplo, significaría que si el número de horas/hombre empleadas es cero, la ecuación de regresión estima un nivel de producción de -15,000 toneladas. Es obvio que, en este caso, esta interpretación es incorrecta. Esto se debe a que los datos observados de X fluctúan entre 757 y 1,195 miles de horas/hombre. En general, la interpretación de las ecuaciones de regresión lineal debe limitarse a puntos entre el rango de valores utilizados para su estimación. El valor numérico de *a* en nuestro ejemplo no es relevante y sólo tiene una interpretación matemática (es el valor de Y cuando X es cero).

El valor del coeficiente de regresión *b*, igual a 0.044, significa que cada unidad adicional de X incrementa el valor de Y en 0.044. En el rango de nuestros datos ($757 \leq X \leq 1{,}195$), podemos decir que, por cada mil horas/hombre adicionales empleadas, predecimos o esperamos 0.044 mil toneladas adicionales de producción de minerales.

En el cuadro 13 se presentan tanto los valores observados de X e Y como los valores estimados de Y, $\hat{Y}$, usando la ecuación de regresión. Además, se calculan las desviaciones de los valores estimados con respecto a los valores observados, ($Y - \hat{Y}$). Estas diferencias son conocidas como los errores de estimación, y su suma es igual a cero, lo que indica que la línea de regresión es

no-sesgada (sus sobreestimaciones son exactamente compensadas por sus subestimaciones). Finalmente, el cuadro 13 muestra los errores al cuadrado que nos permitirán calcular el error estándar de estimación en la próxima sección.

CUADRO 13: CÁLCULO DE $\hat{Y}$ Y SUMA DE ERRORES AL CUADRADO

X	$\hat{Y}$= -15.000 + 0.044X	Y	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1,170	36.48	35.4	-1.08	1.17
1,150	35.60	33.1	-2.50	6.25
1,343	44.09	48.0	3.90	15.21
757	18.31	18.5	0.20	0.04
1,180	36.92	36.4	-0.50	0.25
1,117	34.15	36.0	1.85	3.42
1,180	36.90	34.1	-2.80	7.84
933	26.05	28.1	2.05	4.20
1,195	37.60	39.7	2.10	4.41

3. MEDIDA DEL ERROR ESTÁNDAR DE ESTIMACIÓN

Es deseable disponer de una medida de dispersión de las observaciones individuales con respecto a la línea de regresión, para expresar lo bien o mal que la línea se ajusta a los puntos observados. La medida de dispersión alrededor de la ecuación estimada se llama error estándar de estimación (EEE), se representa por $S_{y.x}$ y se calcula mediante:

$$S_{y.x} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}} \tag{8}$$

Como otras medidas de desviación estándar, el EEE mide la dispersión "promedio" entre los valores observados (Y_i) y los estimados $(\hat{Y}_i)$. De la misma manera, $S_{\bar{X}}$, el error estándar de la media, mide la dispersión de las medias muestrales alrededor de

la media poblacional en el análisis univariable del capítulo IV. Mientras $S_{\bar{x}}$ es una medida de la precisión de la media muestral, $S_{y.x}$ también es una medida del grado de precisión con que la línea de regresión describe la relación entre X e Y. Si las diferencias entre los valores observados y estimados de Y son pequeñas, $S_{y.x}$ será también pequeño, lo que indica que la línea de regresión es una buena descripción de la relación entre X e Y. En el caso extremo, cuando $S_{y.x} = 0$, los valores observados y estimados de Y coinciden y la línea de regresión será una descripción perfecta de la relación, lo que significa que existe una *relación exacta.*

En la ecuación (8), la suma de los cuadrados de los errores es dividida entre (n - 2) para obtener un estimador no-sesgado de la variación alrededor de la línea de regresión «verdadera» de la población ($\sigma_{y.x}$). Se pierde un grado de libertad por cada parámetro estimado, dado que para calcular $S_{y.x}$ previamente se debe calcular $\hat{Y}$, que depende de los parámetros estimados. En este caso de regresión simple estimamos *a* y *b* y, por lo tanto, perdemos dos grados de libertad. En general, en la regresión múltiple perderemos tantos grados de libertad como parámetros estimemos. Así, si tenemos tres variables independientes, estimaremos el intercepto y tres coeficientes de regresión perdiendo cuatro grados de libertad. En tal caso, el denominador de la ecuación (8) será (n - 4).

Usando la ecuación (8) y la colúmna 5 del cuadro 13 podemos calcular el EEE para el ejemplo de la mina "La Hermosa":

$$S_{y.x} = \sqrt{\frac{46.2}{(9-2)}} = 2.569$$

Este valor está expresado en las mismas unidades que Y, es decir, en miles de toneladas. Si cambiamos las unidades de medida de la variable dependiente, el valor numérico del EEE también cambiará. Es por esto que el EEE no es considerado como la medida más conveniente del grado de relación entre variables. Además, el EEE mide cuán pobre es nuestra estima-

ción; pero estamos realmente interesados en saber cuán buenas estimaciones podemos hacer. El coeficiente de correlación es una medida de lo buena que es la línea de regresión y no es afectado por las unidades de medida de la variable dependiente, como se verá en la sección 5.

El EEE permite que se hagan aseveraciones probabilísticas acerca de la confiabilidad de los estimados basados en la línea de regresión ($\hat{Y}_i$). Sabemos que el EEE mide la dispersión alrededor de la línea de regresión; y si suponemos que los puntos observados ($\hat{Y}_i$) se distribuyen normalmente alrededor de la línea de regresión, entonces se podría decir que el intervalo $\hat{Y} \pm S_{y.x}$ incluye el 68% de los puntos muestrales. Del mismo modo, diremos que el intervalo $\hat{Y} \pm 2S_{y.x}$ abarca el 95.5% de los puntos muestrales. También podemos afirmar que existe una probabilidad del 68% de que un estimado hecho con base en la ecuación de regresión, $\hat{Y}_i$, no se desvíe en más de un EEE con respecto al valor verdadero. Este resultado permite establecer intervalos de confianza para los valores estimados.

Debemos notar que se está suponiendo que para cada valor de X en el rango muestral, no importa cuán distante esté de $\overline{X}$, la variable dependiente Y mantiene la distribución normal con el mismo grado de dispersión. Este supuesto se conoce con el nombre de *homoscedasticidad*. Sin embargo, en algunos casos, a medida que se incrementa el valor de la variable independiente, la dispersión de la variable dependiente alrededor de la línea de regresión también aumenta. Es decir que el EEE no es constante para todos los valores de la variable independiente, enfrentándose el problema de *heteroscedasticidad*. Esto conduce a que el método de mínimos cuadrados ordinarios arroje estimaciones sesgadas e ineficientes del EEE (es decir, que sea mayor que el mínimo, y así resulte en pruebas estadísticas incorrectas e intervalos de confianza también incorrectos). Existen métodos alternativos de estimación para salvar este problema, los cuales pueden ser consultados en textos especializados de econometría.

4. USO DEL ERROR ESTÁNDAR DE ESTIMACIÓN PARA PREDICCIÓN

El análisis de regresión tiene tres funciones básicas, que están interrelacionadas. Primero, probar hipótesis sobre la relación que existe entre la variable dependiente y las variables que se supone la explican. Segundo, hacer estimaciones numéricas de los coeficientes de las relaciones formuladas. Finalmente, el análisis de regresión permite predecir valores de la variable dependiente, con el fin de tomar acciones o prevenciones para enfrentar adecuadamente estos valores.

Denotemos X_p al valor de la variable independiente para el período de predicción (p). Sustituyendo este valor en la ecuación de regresión, obtenemos un estimado puntual de la variable dependiente, $\hat{Y}_p$:

$$\hat{Y}_p = a + bX_p \tag{9}$$

Como se vio en el capítulo IV, es preferible establecer un intervalo de confianza para hacer las estimaciones respectivas. Para esto, primero debemos definir la distribución de los errores de predicción, la diferencia entre $\hat{Y}_p$ y el valor real Y, y luego establecer un intervalo de predicción de Y.

Se puede demostrar que la varianza del error de predicción, $(Y_p - \hat{Y}_p)$, de un valor particular de Y cuando $X = X_p$, es igual a:

$$S_{p.x}^2 = S_{y.x}^2 \left[1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right] \tag{10}$$

y el error estándar de predicción será:

$$S_{p.x} = S_{y.x} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum (X_i - \overline{X})^2}} \tag{11}$$

A medida que el número de observaciones, n, se incrementa, el segundo y tercer términos dentro de la raíz cuadrada se hacen más pequeños, y el error estándar de predicción también decrece, acercándose a $S_{y.x}$. De igual manera, mientras más se acerque el valor de X_p a la media, $\bar{X}$, más pequeño será el valor de $S_{p.x}$; y, al contrario, mientras más alejado esté el valor de X_p de $\bar{X}$, mayor será el valor de $S_{p.x}$.

Utilizando el error estándar de predicción, podemos construir un intervalo de confianza de $(1 - \alpha)$ alrededor de la predicción puntual, $\hat{Y}_p$:

$$\hat{Y}_p \pm t_{\alpha/2}\, S_{p.x} \tag{12}$$

El intervalo de predicción tendrá una mayor amplitud cuando el valor de X utilizado para efectuar la proyección esté lejos de la media muestral $\bar{X}$; y será menos amplio a medida que el tamaño muestral usado para la estimación sea mayor.

En nuestro ejemplo de la mina "La Hermosa", podemos predecir el nivel de producción para cierto número de horas/hombre empleadas. Si se planea usar 1,200 miles de horas/hombre el próximo mes ($X_p = 1{,}200$), obtendremos primero una predicción puntual del nivel de producción promedio usando la ecuación de regresión (7):

$$\hat{Y}_p = -15.000 + 0.044\ (1{,}200)$$

$$= 37.8 \text{ miles de toneladas}$$

Luego, podemos establecer un intervalo de confianza del 95% para el nivel de producción esperado usando las expresiones (11) y (12):

$$Y_p \pm t_{\alpha/2} S_{y.x} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum (X_i - \overline{X})^2}}$$

$$37.8 \pm (2.365)\,(2.569) \sqrt{1 + \frac{1}{9} \frac{(1{,}200 - 1{,}114)^2}{230{,}981}}$$

$$37.8 \pm 6.496$$

Podemos afirmar, entonces, con un nivel de confianza del 95%, que si el próximo mes se utilizan 1,200 miles de horas/hombre, la producción estará entre 31.3 y 44,200 toneladas de mineral.

5. COEFICIENTE DE DETERMINACIÓN Y COEFICIENTE DE CORRELACIÓN

Cuanto más cerca estén los valores observados de la línea de regresión –es decir, cuanto más pequeños sean los residuos o errores–, mayor será la variación de la variable dependiente que es explicada por la ecuación de regresión estimada. En esta sección estableceremos una medida de lo buena que es nuestra ecuación de regresión para predecir los valores de Y.

Podemos expresar la variación total en Y como la suma de la variación explicada por la ecuación de regresión, más la residual o no explicada. Para visualizarlo, tomemos cualquier punto de nuestro diagrama de dispersión (gráfico 41) y expresemos la desviación de dicha observación respecto de su media muestral, $(Y_i - \overline{Y})$, como la suma de dos partes:

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_{i}) + (\hat{Y}_i - \overline{Y}) \qquad (13)$$

Llevando a cabo operaciones algebraicas, se demuestra que se cumple la siguiente relación entre la suma de cuadrados de las diferencias de la ecuación (13):

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \qquad (14)$$

Variación total	Variación explicada	Variación residual

Si denotamos la variación total como la suma de cuadrados total (SCT), la variación explicada como la suma de cuadrados de regresión (SCR) y la variación residual como la suma de cuadrados de errores (SCE), tendremos:

$$SCT = SCR + SCE \qquad (15)$$

Esta partición de la suma de cuadrados ayuda en la interpretación de la contribución hecha por X para la explicación de Y. Así, dividiendo ambos lados de la ecuación (15) por SCT, se obtiene:

$$1 = \frac{SCR}{SCT} + \frac{SCE}{SCT} \qquad (16)$$

El *coeficiente de determinación*, R^2, se define como la proporción de la variación total de Y, «explicada» por la regresión de Y en X:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \overline{Y})^2} \qquad (17)$$

Esta medida nos indica cuán buena es la línea de regresión para explicar o predecir la variable dependiente, dado que mide la razón entre la variación explicada (SCR) y la variación total (SCT).

En nuestro ejemplo particular de la mina «La Hermosa», tenemos:

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = 0.91$$

Por lo tanto, podemos concluir que el 91% de la variación en la producción de minerales de los últimos nueve meses está relacionada, o explicada, por la variación en las horas/hombre empleadas.

Dado que hemos calculado R^2 basándonos en información muestral, este es un estadígrafo; pero es un estimador sesgado del parámetro ρ^2 (rho cuadrado). El parámetro ρ^2 se obtendría si se calculase el coeficiente de determinación con todos los puntos de la población. El valor esperado de R es mayor que ρ^2. Para corregir este sesgo, reemplazamos las sumas de cuadrados en la ecuación (17) por $S_{y.x}^2$ y S_y^2. Esto nos da:

$$\overline{R}^2 = 1 - \frac{S_{y.x}^2}{S_y^2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2/(n-2)}{\sum(Y_i - \overline{Y})^2/(n-1)} \tag{18}$$

$\overline{R}^2$ es un estimador no-sesgado de ρ^2, y se le conoce como el *coeficiente de determinación ajustado* por los grados de libertad. Los estadígrafos $S_{y.x}^2$ y S_y^2 son estimadores no-sesgados de $\sigma_{y.x}^2$ y σ_y^2, respectivamente.

En nuestro ejemplo, tenemos:

$$\overline{R}^2 = 0.89$$

Si comparamos este valor con el de R^2, vemos el sesgo positivo de R en la estimación de ρ^2. En la muestra de nueve observaciones, el número de horas/hombre empleadas explica el 91% de la variación en los niveles de producción de minerales. Pero no podemos usar ese valor para estimar la proporción de la

variación de la producción que es explicada por el número de horas/hombre empleadas para toda la «población» o universo. El estimado correcto de esta proporción es $\overline{R}^2 = 0.89$. Considerando este ajuste, diremos que el 89% de la variación total de la producción de minerales es explicada por el número de horas/hombre empleadas para su producción.

Consideremos ahora los posibles valores numéricos que puede tomar el coeficiente de determinación. Si existe una relación exacta entre X e Y, entonces los valores observados y los valores estimados son idénticos –por tanto, $\sum (Y_i - \hat{Y}_i)^2 = 0$ –, y la variación explicada es igual a la variación total. En términos de las ecuaciones (17) y (18), tendremos $R^2 = 1 - 0 = 1$, y $\overline{R}^2 = 1$, respectivamente. Este es el valor más alto que el coeficiente de determinación puede tomar, indicando una correlación perfecta entre las dos variables ya que toda la variación de Y está explicada por la variación de X. En el otro extremo, si X e Y no están relacionadas, la variación explicada es cero, y por lo tanto $R^2 = 0$. En consecuencia, el rango de variación del coeficiente de determinación está entre 0 y 1.

El *coeficiente de correlación, r,* es la raíz cuadrada de R^2. Entonces su valor numérico está entre -1 y 1. La virtud principal de esta medida es que su signo indica la dirección de la relación: r es positivo cuando X e Y están relacionados directamente, y es negativo cuando X e Y están relacionados inversamente. Entonces, el signo de r es realmente el signo de la pendiente de la ecuación de regresión; es decir, el signo del coeficiente de regresión b.

En nuestro ejemplo, $R^2 = 0.91$, y, por tanto, $r = \sqrt{0.91} = 0.95$, positivo, ya que X e Y están relacionadas directamente, como lo indica el signo de b = 0.04

Un comentario final respecto a la interpretación de r y R^2. Tanto r como R^2 miden el grado de relación líneal entre X e Y: ambos toman el valor 1 cuando la correlación es perfecta y el valor cero cuando no hay ninguna correlación. A excepción de estos casos extremos, R^2 es el más significativo, porque es el que

mide la proporción de la variación en Y explicada por la variación en X.

6. PRUEBA DE SIGNIFICACIÓN DEL COEFICIENTE DE DETERMINACIÓN

Obsérvese que el coeficiente de correlación muestral *r* es un estimador del coeficiente de correlación poblacional. Como vimos en el capítulo anterior, es posible que un estadígrafo (r) difiera del parámetro poblacional ($\int$). En nuestro ejemplo obtuvimos un $R^2 = 0.91$; es decir, que el 91% de la variación de Y es explicada por X. La pregunta relevante es: ¿existe realmente una relación entre el nivel de producción y las horas/hombre empleadas, o este valor de R^2 se debe únicamente a los errores muestrales? Dado que nuestra preocupación principal es establecer si existe o no correlación entre X e Y, podremos usar el procedimiento de contraste de hipótesis desarrollado en el capítulo V. Tendremos una hipótesis nula de $\int = 0$, que expresa que no existe relación entre X e Y. Siguiendo el procedimiento de prueba de hipótesis, tendremos:

(i) $H_o : \int = 0$
$H_1 : \int \neq 0$

(ii) Usando un *nivel de significación* (α) de 0.05 para una *prueba de dos colas* (dada la igualdad estricta en H_o), el valor crítico de t con *siete grados de libertad* será ± 2.365. Note que perdemos dos grados de libertad porque hay dos parámetros en la ecuación de regresión.

(iii) La *regla de decisión* será:
Rechazar H_o si $|t| > 2.365$. La distribución muestral de r se ha aproximado a una distribución t de Student, dado que $\int$ varía entre -1 y +1, con una media igual al parámetro de la población $\int$, y una desviación estándar estimada

$S_r = \sqrt{(1 - R^2)/(n\ 2)}$. Note que $1 - R^2$ es la proporción que no podemos explicar con la ecuación de regresión. Matemáticamente:

$$r \sim t\,(\rho, S_r) \qquad (19)$$

Luego, el valor calculado de t será:

$$t = \frac{r - \rho}{S_r} \qquad (20)$$

dado que nuestra hipótesis nula es $\rho = 0$,

$$t = \frac{r}{\sqrt{(1 - R^2)/(n - 2)}} \qquad (21)$$

En nuestro ejemplo, tenemos: $r = 0.95$ y $R^2 = 0.91$; entonces:

$$t = \frac{0.95}{\sqrt{(1 - 0.91)/7}} = 8.38$$

Dado que el t calculado es mayor que el t crítico ($8.38 > 2.365$), no podemos aceptar la hipótesis nula, concluyendo que el coeficiente de correlación es significativamente diferente de cero, y por tanto hay una relación entre los niveles de producción de minerales y las horas/hombre empleadas para su producción.

7. PRUEBA DE SIGNIFICACIÓN DEL COEFICIENTE DE REGRESIÓN (b)

El coeficiente de regresión, *b*, que es la pendiente de la ecuación de regresión, mide el cambio esperado en Y ante un cambio unitario en X. Este estadígrafo es un estimador del coeficiente

de regresión poblacional, . Si X no proporcionara información para predecir Y, esto implicaría que $\beta = 0$. Entonces podemos someter a prueba la hipótesis de que $\beta \neq 0$ contra la alternativa de $\beta = 0$; es decir, que X ayuda a explicar a Y.

Dado que hemos supuesto que todos los posibles valores de Y para un determinado valor de X se distribuyen normalmente con una media igual a $\hat{Y}$ y una desviación estándar igual a $S_{y.x}$, se puede probar que los coeficientes de regresión tienen una distribución normal con los siguientes parámetros:

$$E(b) = \beta; \quad \sigma_b^2 = \sigma^2/\sum(X - \bar{X})^2 \tag{22}$$

Es decir:

$$b \sim N(\beta, \sigma_b) \tag{23}$$

Para desarrollar el procedimiento de contraste de hipótesis calculamos el estadístico z, que tiene una distribución normal estandarizada para muestras repetitivas:

$$z = \frac{b - \beta}{\sigma_b} \tag{24}$$

Como usualmente no se conoce σ, debemos primero estimarla, y, subsecuentemente, calcular σ_b por $S_{y.x}$ en la ecuación (22). Así, obtenemos:

$$S_b = \frac{S_{y.x}}{\sqrt{\sum(X - \bar{X})^2}} \tag{25}$$

Y sustituyendo S_b por σ_b en la ecuación (24), el estadístico z se convierte en el estadístico t:

$$t = \frac{b - \beta}{S_b} \tag{26}$$

Este estadístico tiene una distribución t de Student con (n-2) grados de libertad, que son los mismos asociados a $S_{y.x}$.

Si deseamos probar la hipótesis de que β es igual a algún valor preestablecido, utilizaremos el procedimiento estándar de contraste de hipótesis antes explicado. Usualmente se desea probar si el coeficiente de regresión es significativamente diferente de cero ($\beta \neq 0$), y entonces, siguiendo el procedimiento de prueba de hipótesis, tendremos:

(i) $H_0 : \beta = 0$
$H_1 : \beta \neq 0$

(ii) Usando un $\alpha = 0.05$ para una prueba de dos colas, el valor crítico de t, con siete grados de libertad, será $\pm$ 2.365.

(iii) La regla de decisión será:
Rechazar H_0 si $|t| > 2.365$
Donde: $t = b/S_b$ de acuerdo con la ecuación (28), con $\beta = 0$.
En nuestro ejemplo de la mina "La Hermosa", tenemos: $b = 0.044$, $S_{y.x} = 2.569$ y $\sum (X - \overline{X})^2 = 476$; entonces:

$$t = \frac{0.044}{2.569/\sqrt{476}} = 8.31$$

Como $t = 8.31 > 2.365$, rechazamos la hipótesis nula y concluimos que el coeficiente de regresión es significativamente diferente de cero, y por tanto hay evidencia que indica que las horas/hombre empleadas proporcionan información relevante para las predicciones de los niveles de producción de minerales.

8. CONDICIONES PARA EL USO DEL MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS

Para estimar el modelo de regresión lineal, ecuación (4), usando el método de mínimos cuadrados ordinarios (MCO), se establecieron los siguientes supuestos:

1. La ecuación de regresión está correctamente especificada. Es decir, existe en efecto una relación lineal que fue determinada del diagrama de dispersión. Además, significa que X es la única variable relevante para explicar Y. El modelo de regresión múltiple presentado en la sección 10 levanta este último supuesto.

2. La variable independiente no es aleatoria; X es fija en diferentes muestras.

3. Los valores de Y son independientes unos de otros. Si los valores de Y están relacionados por otra relación que la del modelo de regresión lineal, diremos que están autocorrelacionados. Una prueba estándar de autocorrelación es la prueba de Durbin Watson, que se presenta en los textos de econometría.

4. Para cada valor de X, los correspondientes valores de Y se distribuyen normalmente alrededor del valor estimado correspondiente, $\hat{Y}$, con una desviación estándar constante, independiente del valor de X. Matemáticamente:

$$(Y/X) \sim N(\hat{Y}, S_{y.x}) \qquad (27)$$

El que la desviación estándar sea constante e igual al EEE, se conoce como *homocedasticidad*. Si la desviación estándar de Y variara para diferentes valores de X, tendríamos un problema de *heteroscedasticidad*.

9. ANÁLISIS DE CORRELACIÓN LINEAL

En el análisis de regresión se trata a la variable Y como una variable dependiente que está explicada por X, la variable independiente. En cambio, el análisis de correlación mide la *interdependencia* entre dos variables sin afirmar que una de ellas sea dependiente de la otra. Es un análisis simétrico, en el sentido de que las dos variables son tratadas en forma idéntica. En la fórmula del coeficiente de correlación se puede intercambiar X e Y y obtener el mismo resultado.

El diagrama de dispersión puede también ser usado para indicar el grado de asociación o *correlación* entre dos variables. El gráfico 42 muestra diferentes tipos de relación entre dos variables.

Cuando se establece que la relación entre dos variables es lineal, la medida de correlación comúnmente usada es el llamado *coeficiente de correlación simple o de Pearson* entre X e Y, y se le denota por r.

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} \qquad (28)$$

Podemos comprobar, con esta fórmula, que el concepto de correlación es simétrico; es posible intercambiar las X con las Y sin cambiar la fórmula. Es decir, la correlación entre X e Y es la misma que la correlación entre Y y X. En contraste, la regresión de Y en X no es equivalente a la regresión de X en Y.

El estadígrafo r se calcula con base en los datos muestrales. Este es el estimado del verdadero coeficiente de correlación poblacional, $\int$, definido por:

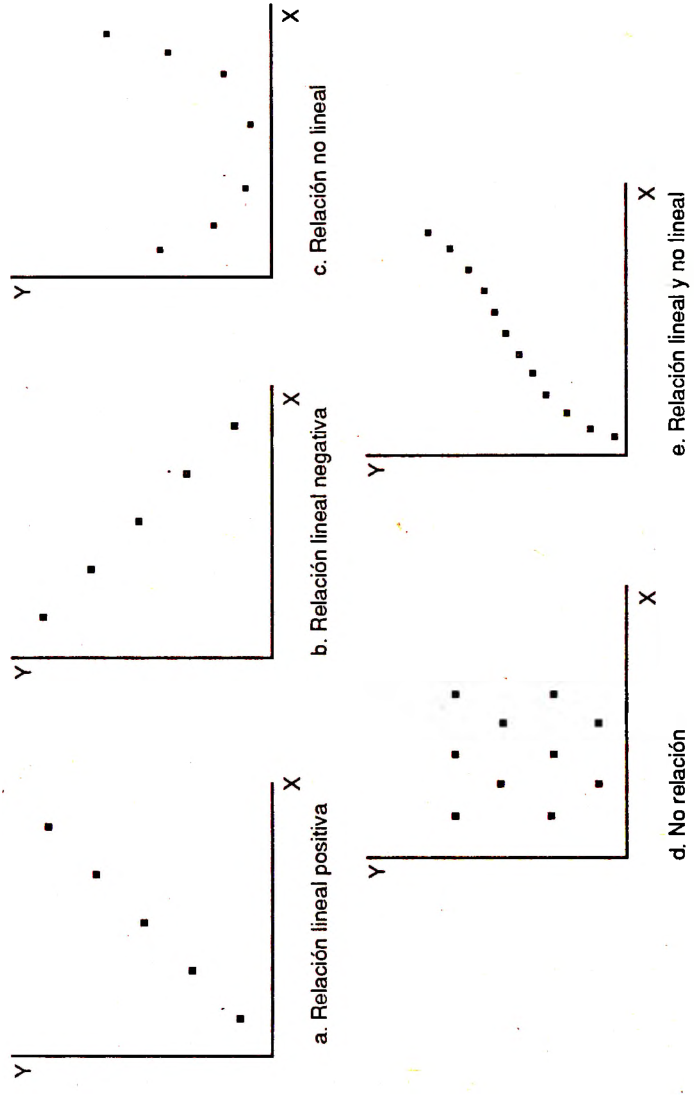
$$\int = \frac{\sum (X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum (X_i - \mu_x)^2 \sum (Y_i - \mu_y)^2}} \qquad (29)$$

Si se define:

$$\text{cov}(X, Y) = \frac{\sum (X_i - \mu_x)(Y_i - \mu_y)}{N}$$

$$\sigma_x = \sqrt{\sum (X_i - \mu_x)^2 / N} \; ; \quad \sigma_y = \sqrt{\sum (Y_i - \mu_y) / N}$$

Gráfico 42: Ilustración del tipo de relaciones usando los diagramas de dispersión



entonces podemos reescribir la ecuación (29) como:

$$\int = \frac{\text{cov}\,(X, Y)}{\sigma_x\,\sigma_y} \qquad (30)$$

donde N es el tamaño poblacional.

μ_x es la media poblacional de X.

μ_y es la media poblacional de Y.

cov (X,Y) es la covarianza poblacional entre X e Y.

σ_x es la desviación estándar poblacional de X.

σ_y es la desviación estándar poblacional de Y.

Ello quiere decir que el coeficiente de correlación es el cociente entre la covarianza de X e Y y el producto de las desviaciones estándar de X e Y. El signo de $\int$ depende del signo de la cov(X,Y). Por inspección podemos establecer el signo de la covarianza de X e Y. Cuando exista una relación positiva entre X e Y, los valores altos de X estarán asociados con valores altos de Y (ver gráfico 42). Cuando tanto los valores de X como los de Y están por encima de sus medias respectivas, las diferencias $(X_i - \mu_x)$ e $(Y_i - \mu_y)$ serán positivas, y el producto $(X_i - \mu_x)(Y_i - \mu_y)$ será también positivo. De igual manera, esperamos que valores bajos de X estén asociados con valores bajos de Y. Cuando los valores de X e Y están por debajo de sus medias respectivas, las diferencias $(X_i - \mu_x)$ e $(Y_i - \mu_y)$ serán negativas, pero su producto $(X_i - \mu_x)\ (Y_i - \mu_y)$ será otra vez positivo. Por lo tanto, si la relación es positiva el numerador de la ecuación (29) será positivo, al ser la sumatoria de productos donde los elementos positivos predominan, y $\int$ será positivo.

Por otro lado, si la relación es negativa esperamos que valores altos de X estén asociados con valores bajos de Y, y viceversa,

de tal manera que las diferencias ($X_i - \mu_x$) e ($Y_i - \mu_y$) serán generalmente de signo contrario (ver gráfico 42 (b)). Luego, el producto ($X_i - \mu_x$) ($Y_i - \mu_y$) será negativo, y ambos términos predominarán en la sumatoria, haciendo que el numerador de la ecuación (29) sea negativo, por lo que el coeficiente de correlación r también será negativo.

Si las variables aleatorias X e Y son independientes (ver gráfico 42 (d)), entonces la cov (X,Y) tenderá a cero, al igual que $\int$. En este caso, si tomamos una muestra aleatoria del universo de pares (X, Y), el estadígrafo r tendrá un valor esperado de cero. Si se quiere probar que $\int$ es cero o, alternativamente, que $\int$ es significativamente diferente de cero, es necesario determinar la distribución muestral del estadígrafo r para definir el estadístico adecuado (t o z). Podemos usar el procedimiento de prueba de hipótesis de la sección 6, dado que el valor numérico del coeficiente de correlación de Pearson coincide con el valor del coeficiente de correlación del análisis de regresión en el caso en que sólo existe una variable independiente.

Si calculamos el coeficiente de correlación lineal de Pearson para los datos del cuadro 12, encontramos que r = 0.948, igual al coeficiente de correlación del análisis de regresión simple salvo por los errores de redondeo en los cálculos. Como se dijo anteriormente, la diferencia entre el análisis de correlación y el análisis de regresión simple es la manera como son tratadas las variables incluidas en el análisis. El coeficiente de correlación es una medida de cómo varían ambas variables y es, por tanto, un enfoque simétrico. El análisis de regresión tiene un enfoque diferente: considera a una de las variables como conocida, llamándola variable independiente. Desarrolla una ecuación para predecir los valores de la otra variable, la variable dependiente, tomando como base los valores de la primera. Entonces, el análisis de regresión es asimétrico: la ecuación de regresión que predice los valores de Y con base en los valores de X es totalmente diferente de la ecuación de

regresión que predice los valores de X con base en los valores de Y, aun cuando las ecuaciones sean estimadas usando los mismos datos.

10. ANÁLISIS DE REGRESIÓN MÚLTIPLE

Las técnicas del análisis de regresión simple y del análisis de correlación discutidas en las secciones anteriores consideran sólo dos variables: X e Y. Pero la complejidad de la mayoría de los problemas del mundo real requiere que se considere más de dos variables en el análisis de dichos problemas. El coeficiente de correlación lineal de Pearson es una medida de cómo varían dos variables conjuntamente, estableciendo su correlación "bruta". Con este concepto no es posible medir cuánto de esta relación es asociada con una tercera o cuarta variable que podría estar también relacionada a las dos primeras. Las variables X e Y podrían variar conjuntamente, pero también Z podría variar con ellas. Algunas veces nos gustaría mantener a Z constante, y observar el efecto que tiene X en Y que no es atribuible a Z; es decir, la relación "neta" entre X e Y. El análisis de regresión múltiple permite cuantificar estas relaciones netas.

Si R^2 resulta muy pequeño en el análisis de regresión simple, la ecuación de regresión estimada será de utilidad cuestionable. Una mayor explicación de las variaciones de la variable dependiente (*i.e.*, un mayor R^2) quizá podría lograrse construyendo una ecuación de regresión para Y basada no sólo en una variable sino en un conjunto de variables independientes (X_1, X_2..., X_k). Así, definimos una ecuación de regresión lineal múltiple que tiene la forma general:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \qquad (31)$$

Esta ecuación de regresión lineal múltiple predice el valor de la variable dependiente con base en la información de dos o más variables independientes ($X_1, X_2 \ldots, X_k$). La ecuación (31) es *lineal*, dado que sus términos sólo contienen constantes y variables independientes elevadas a la primera potencia; no incluye raíces cuadradas, cuadrados, productos cruzados de variables, etcétera.

En el caso de que k sea 1, se tiene la ecuación de regresión simple, y la ecuación se grafica como una línea recta. Cuando $k = 2$, existen dos variables independientes e $\hat{Y}$ se puede graficar como un plano. Cuando se definen tres o más variables, $K \geq 3$, la representación gráfica de la ecuación es imposible, pero a la ecuación estimada Y se le conoce como un hiperplano lineal.

La estimación de a, b_1, b_2,... es compleja y está fuera de los alcances de este libro. Nos limitaremos a decir que estos coeficientes se obtienen a través de minimizar la suma de los errores de estimación al cuadrado, al igual que en la regresión simple. Matemáticamente:

$$\text{Min}\sum(Y_i - \hat{Y}_{i})^2 = \text{Min}(Y_i - a - b_i X_{1i} - b_2 X_{2i} - \ldots - b_k Y_{ki})^2$$

Al minimizar esta función se obtiene un sistema de (k + 1) ecuaciones en (k + 1) incógnitas cuya solución se facilita tremendamente con la ayuda de un programa de computadora. En esta sección discutiremos la interpretación de los resultados que nos ofrece uno de estos paquetes.

El valor $\hat{Y}$ en el lado izquierdo de la ecuación (31) es el valor estimado de Y. El primer término en el lado derecho de dicha ecuación es la constante a, que indica el valor de Y cuando todas las variables independientes son cero. La interpretación significativa del valor de «a» en el contexto del análisis depende de si los datos relacionados a las variables independientes X pueden tomar el valor de cero, todas a la vez. En caso contrario, la interpretación del valor de «a» es irrelevante.

El segundo término, $b_1 X_1$, expresa que por cada unidad de incremento de X_1, mientras $X_2...X_k$ permanecen constantes, el valor de Y aumentará en b_1. Lo mismo ocurre con el resto de términos de la ecuación de regresión múltiple.

Los coeficientes b_1, $b_2...b_k$ son conocidos como los *coeficientes de regresión parcial,* dado que nos expresan el cambio en Y debido a un cambio unitario en una de las X, mientras las otras variables independientes permanecen constantes. Es decir, que los coeficientes de regresión miden el *efecto neto* entre una variable independiente y la variable dependiente.

Además de los supuestos establecidos en el modelo de regresión simple, hay un supuesto adicional en el modelo de regresión múltiple: la no existencia de una relación lineal exacta entre las variables independientes X_1, $X_2...X_k$. De no cumplirse este supuesto, no se podrá resolver el sistema de ecuaciones para estimar los coeficientes de regresión parcial, y, por tanto, ningún otro estadígrafo.

Por otro lado, se dan casos en los que dos o más variables independientes de la ecuación de regresión múltiple están altamente correlacionadas, haciendo difícil o imposible aislar sus efectos individuales sobre la variable dependiente. Esto se conoce como el problema de *multicolinealidad.* En casos en los que la multicolinealidad está bien definida, los coeficientes estimados (b_i) pueden resultar estadísticamente insignificantes, o, incluso, tener un signo contrario al esperado. En otros casos la detección de la multicolinealidad no es tan obvia. Los coeficientes de correlación simples entre las variables independientes se usan como una medida de multicolinealidad. Se debe notar que puede presentarse multicolinealidad aun si los coeficientes de correlación simple son relativamente bajos (*i.e.*, $r < 0.5$).

La multicolinealidad puede corregirse o reducirse ampliando el tamaño de la muestra, omitiendo una de las variables altamente colineales, o transformando la relación funcional.

Para establecer si un coeficiente de regresión parcial es estadísticamente insignificante, necesitamos especificar procedimientos de prueba de hipótesis, como lo hicimos en la sección 7 para el modelo de regresión simple. Bajo la hipótesis nula de que el parámetro poblacional sea igual a cero, $\beta_i = 0$, y dado un nivel de significación, se podrá establecer el valor crítico de "t" con [n - (k + 1)] grados de libertad. La decisión de aceptar o rechazar la hipótesis nula se tomará comparando el valor crítico de t con el valor calculado de t:

$$t = \frac{b_i}{S_{bi}} \qquad (32)$$

Estas pruebas suponen que los datos muestrales son observaciones independientes de una población que tiene la relación:

$$E\,(Y/X_1, X_2, \ldots X_n) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k \qquad (33)$$

Además, se presume que los valores de Y se distribuyen normalmente alrededor de la ecuación de regresión, y que la desviación estándar ($S_{y.x1,\ x2...xk}$) es constante para todos los valores de X_1, $X_2...X_k$ (homoscedasticidad), como en el modelo de regresión simple.

11. EL ANÁLISIS DE REGRESIÓN Y LA COMPUTADORA

En este capítulo se han discutido las técnicas del análisis de regresión. Para ilustrar el análisis de regresión simple se presentó el caso de la mina "La Hermosa" considerando una muestra pequeña de nueve observaciones para simplificar los cálculos. Sin embargo, al usar esta muestra pequeña sacrificamos la precisión de las estimaciones de la ecuación de regresión. Por otro lado, en la discusión del análisis de regresión múltiple no

se presentó ningún ejemplo, para evitar hacer cálculos tediosos. La exposición completa hubiera implicado la manipulación de matrices.

En general, en el proceso de toma de decisiones nos enfrentamos a problemas complejos que involucran a más de una variable independiente. Asimismo, los modelos de regresión que explican el comportamiento y relación entre estas variables requieren de un número significativo de observaciones para estimar adecuadamente los estadígrafos correspondientes. Para ayudarnos a estimar estos modelos más complejos se han desarrollado programas de computadora que facilitan enormemente los cálculos numéricos. Además, algunos programas también permiten estimar regresiones curvilíneas para describir relaciones no lineales.

Los programas estadísticos para microcomputadora de mayor uso en nuestro medio son: TSP (Time Series Package), Statpac, Microstat y RATS (Regression Analysis Time Series).

Consideremos el siguiente ejemplo para ilustrar los resultados que nos ofrece el TSP. "Manufacturas Metálicas Eléctricas S.A." ha visto descender en los últimos meses las ventas de sus transformadores para fluorescentes "Ajax". El Gerente Comercial consideró que la retracción de las ventas podría ser explicada por una serie de variables independientes: el precio de "Ajax" en dólares (PAJDOL), el precio relativo de "Ajax" versus precios del bien sustituto (PAJSUS), el precio relativo de "Ajax" versus el precio promedio de la competencia (PCOMP), el producto bruto interno del sector construcción (PBI), el sueldo real promedio per cápita a nivel nacional (SUELD1) y el consumo de energía eléctrica en Lima Metropolitana (GWH). Las últimas tres variables macroeconómicas se incluyen como indicativas del nivel de actividad económica. Las observaciones de los últimos 54 meses para estas variables se presentan en el cuadro 14.

CUADRO 14: SERIES DE TIEMPO DE LAS VARIABLES DEL PROBLEMA DE "AJAX"

Obs.	VAJAX	PAJDOL	PAJSUS	PCOMP	PBI	SUELD1	GWH
1	12,760.00	2.892805	5.344547	1.015464	16.60000	19.19534	172.6000
2	10,120.00	2.936339	6.006511	1.029767	14.40000	19.19540	173.4000
3	14,060.00	3.060427	7.007596	1.041532	15.60000	19.07419	168.0000
4	17,340.00	3.286624	8.399348	0.982857	14.70000	18.30157	170.2000
5	13,620.00	3.212465	9.229517	1.015522	14.20000	20.47290	170.4000
6	9,400.00	3.223805	6.680563	1.022872	13.10000	17.62635	175.1000
7	24,460.00	3.364985	6.822078	1.007105	14.80000	18.27670	171.2000
8	8,140.00	2.611399	7.057725	1.007105	12.30000	17.43227	169.9000
9	7,720.00	2.590520	8.966199	1.007105	13.60000	19.82591	180.5000
10	13,730.00	2.590520	8.966199	1.007105	16.00000	19.91156	177.6000
11	14,020.00	2.597938	8.966199	1.007105	16.10000	20.21211	177.3000
12	23,440.00	2.605399	8.113039	1.007105	16.90000	21.75841	175.5000
13	27,702.00	2.600917	7.054433	1.007105	17.50000	21.12756	174.5000
14	2,420.00	2.603904	6.526619	1.007105	16.10000	15.37176	171.8000
15	36,542.00	2.723624	7.648953	1.054618	14.80000	22.63997	171.3000
16	37,924.00	2.723624	7.057950	1.001265	17.00000	22.64557	168.2000
17	57,970.00	2.723624	7.057950	1.001265	17.10000	22.40593	175.3000
18	32,333.00	2.723624	7.057950	1.001265	17.30000	23.75346	180.3000
19	37,680.00	2.715838	7.057950	1.001265	18.60000	23.98116	181.9000
20	14,220.00	2.685133	7.057950	1.001265	19.30000	23.58270	186.2000
21	41,954.00	2.871094	7.644874	1.084528	19.90000	23.82505	196.4000
22	30,910.00	2.778078	7.644874	1.084528	20.60000	26.02639	193.4000
23	38,850.00	2.699371	7.644874	1.084528	20.20000	26.07642	203.4000
24	23,265.00	2.539487	7.644874	1.084528	22.80000	25.95599	197.3000
25	52,950.00	3.266856	8.575496	1.460582	21.50000	25.16360	200.5000
26	59,018.00	3.391581	8.170991	1.217323	18.30000	24.50491	194.7000
27	69,304.00	3.355820	8.330189	0.912543	21.80000	24.39918	197.0000
28	81,523.00	3.412895	9.425708	1.032554	20.60000	26.66622	208.0000
29	66,721.00	2.934288	9.425708	1.032554	19.50000	25.52248	198.1000
30	55,038.00	2.648443	8.566988	1.032554	21.30000	24.94568	199.3000
31	41,390.00	1.972119	7.634193	1.032554	20.90000	27.21308	200.0000
32	47,620.00	1.700638	7.634193	1.032554	23.30000	26.75036	210.3000
33	54,438.00	1.537115	7.414657	1.032554	22.90000	26.33408	209.8000
34	39,194.00	1.545333	8.577243	1.197778	23.90000	25.72193	221.4000
35	101,016.00	1.878358	9.793774	1.625759	21.20000	31.00936	225.2000

(viene de la página anterior)

36	53,700.00	1.382967	9.793774	1.102787	21.30000	28.79847	206.5000
37	80,059.00	1.854105	9.604144	1.459804	20.50000	26.07893	212.8000
38	68,786.00	1.726863	9.604144	1.459804	22.70000	25.38896	212.3000
39	67,300.00	2.077170	8.921393	1.658356	25.50000	25.53737	201.8000
40	43,150.00	2.437647	15.111830	2.026846	22.20000	23.88988	229.1000
41	62,624.00	2.233293	9.521573	2.026846	21.60000	23.22402	223.3000
42	50,436.00	2.171781	9.521573	2.026846	21.60000	22.61408	210.4000
43	20,414.00	2.832871	15.115400	2.533570	19.80000	24.05287	218.2000
44	70,669.00	2.123942	15.115400	2.533570	20.30000	22.07240	224.7000
45	5,615.00	7.454031	15.579410	0.931387	20.10000	16.72450	226.4000
46	9,535.00	4.992863	14.003860	1.111300	18.40000	13.49993	225.7000
47	19,455.00	3.686318	8.309022	1.111300	17.90000	16.94352	230.0000
48	15,294.00	1.792127	8.396826	1.389059	15.70000	14.02946	212.2000
49	10,550.00	3.344213	9.761396	1.282472	16.10000	13.15361	190.2000
50	10,453.00	4.301230	9.761396	1.282472	13.20000	11.46279	
51	21,405.00	3.634314	7.838748	1.282472	14.70000	11.00149	171.0000
52	12,250.00	3.605736	10.780510	1.763764	13.80000	9.322459	170.8000
53	19,537.00	2.997474	7.927668	2.083726	15.20000	8.908254	185.3000
54	22,486.00	3.796083	6.959040	2.492515	15.90000	9.190202	184.5000
55	25,586.00	4.519633	5.722995	1.869670	NA	NA	182.3000
56	NA	NA	5.413012	1.110611	NA	NA	NA

donde

VAJAX	Venta de transformadores "Ajax"
PAJDOL	Precio de transformadores "Ajax" en dólares.
PAJSUS	Precio relativo de transformadores "Ajax" versus precio del bien sustituto.
PCOMP	Precio relativo de transformadores "Ajax" versus precio promedio de la competencia.
PBI	Producto bruto interno del sector construcción, en intis de 1979 (*Fuente:* BCR).
SUELD1	Sueldo promedio per cápita a nivel nacional, en intis de 1979 (*Fuente:* BCR).
GWH	Consumo de energía eléctrica en Lima Metropolitana, en gigio watt/hora.

Nota: La información es mensual, desde enero de 1985 hasta junio de 1989.

Se planteó el modelo de regresión múltiple incluyendo todas las variables explicativas consideradas por el Gerente Comercial. Usando el TSP se obtuvieron los resultados que se muestran en el cuadro 15. Del análisis de estos resultados vemos que a pesar de obtener un nivel de explicación relativamente alto (R^2 = 0.64, R^2=0.59), algunas variables no resultan relevantes para explicar las ventas de "Ajax", dado que sus estadísticos t (T-Stat) son más pequeños que los t críticos. Tomemos, por ejemplo, el consumo de energía eléctrica (GWH). El Gerente Comercial piensa que es una variable crucial para explicar las ventas de "Ajax", pero al estimar la ecuación de regresión el TSP nos indica que esta variable no es significativa (su estadístico es

CUADRO 15: RESULTADOS DEL TSP CON SEIS VARIABLES INDEPENDIENTES

SMPL 1-54
54 Observations
LS//Dependent Variable is VAJAX

Variable	Coefficient	Std. Error	T-Stat.	2-Tail Sig.
C	-91982.5	28400.448	-3.2387689	0.002
PBI	731.2557	1378.2858	0.5305545	0.598
SUELD1	2942.16	768.40786	3.8289037	0.000
GWH	207.7892	211.28594	0.9834504	0.330
PCOMP	19321	7095.0463	2.7231671	0.009
PAJDOL	1694.118	3324.466	0.5095909	0.613
PAJSUS	-2112.95	1537.5555	-1.3742277	0.176

R-squared	0,636687	Mean of dependent var	34853.89
Adjusted R-squared	0,590306	S.D. of dependent var	23785.39
S.E. of regression	15.224,40	Sum of squared resid	1.09D+10
Durbin-Watson stat	1,408584	F-statistic	13,72748
Log likelihood	-592,9294		

0.98 menor que el t crítico). Esta discrepancia nos hace pensar en la existencia de un problema de multicolinealidad. Para confirmarlo, se calcularon las covarianzas y coeficientes de correlación simple, siempre con la ayuda del TSP. Los resultados que se presentan en el cuadro 16 indican claramente que existe una alta correlación entre el PBI y GWH; también se observa una alta correlación entre el PBI y SUELD1. En ambos casos el coeficiente de correlación es mayor que 0.72.

CUADRO 16: COVARIANZAS Y COEFICIENTES DE CORRELACIÓN SIMPLE ENTRE LAS VARIABLES DEL PROBLEMA DE "AJAX"

SMPL 1-54
51 Observations

Series	Mean	S.D.	Maximum	Minimum
PBI	18.35556	3.27343	25.500000	12.300000
SUELD1	21.162930	5.313335	31.009360	8.908254
GWH	193.8667	19.63791	230.000000	168.000000
VAJAX	34853.89	23785.39	101,016.000000	2,420.000000
PCOMP	1.272817	0.431324	2.533570	0.912544
PAJDOL	2.8416670	0.9476750	7.454031	1.382967
PAJSUS	8.774723	2.297354	15.579410	5.344547

	Covariance	Correlation
PBI, PBI	10.51691	1.0000000
PBI, SUELD1	12.30635	0.7209029
PBI, GWH	45.745	0.7250420
PBI, VAJAX	52381.95	0.6854662
PBI, PCOMP	0.2523	0.1820660
PBI, PAJDOL	-1.02379	-0.3362524
PBI, PAJSUS	1.969492	0.2668340
SUELD1, SUELD1	27.70872	1.0000000

(sigue)

(viene de la página anterior)

SUELD1, GWH	37.49167	0.3660920
SUELD1, VAJAX	86887.65	0.7004843
SUELD1, PCOMP	-0.37398	-0.1662611
SUELD1, PAJDOL	-2.62092	-0.5303282
SUELD1, PAJSUS	-0.11379	-0.0094982
GWH, GWH	378.5059	1.0000000
GWH, VAJAX	225660.6	0.4922300
GWH, PCOMP	3.34119	0.4019019
GWH, PAJDOL	-1.13081	-0.0619091
GWH, PAJSUS	28.10577	0.6347302
VAJAX, VAJAX	555268214	1.0000000
VAJAX, PCOMP	2201.292	0.2186157
VAJAX, PAJDOL	-9533.64	-0,4309301
VAJAX, PAJSUS	6474.35	0.1207191
PCOMP, PCOMP	0.182595	1.0000000
PCOMP, PAJDOL	-0.04844	-0.1207398
PCOMP, PAJSUS	0.484172	0.4978357
PAJDOL, PAJDOL	0.881457	1.0000000
PAJDOL, PAJSUS	0.66081	0.3092481
PAJSUS, PAJSUS	5.180097	1.0000000

Como dijimos en la sección 9, una manera de corregir el problema de multicolinealidad es omitiendo alguna de las variables altamente colineales. Luego de plantear diferentes especificaciones para el modelo de regresión considerando la información del Gerente Comercial, se definió una ecuación de regresión final con sólo dos variables explicativas: el consumo de energía (GWH) y el sueldo promedio per cápita (SUELD1). Los resultados de la estimación de este modelo, que se presentan en el cuadro 17, muestran que ambas variables son significativas para explicar la variación de las ventas de "Ajax", ya que sus respectivos "t" son superiores al t crítico, 1.96. El R^2 (ajustado) ha disminuido con respecto a la ecuación del cuadro 15 de 59% a 54%, pero ya no existe el problema de multicolinealidad.

CUADRO 17: ECUACIÓN DE REGRESIÓN CON DOS VARIABLES INDEPENDIENTES

SMPL 1-54
54 Observations
LS// Dependent Variable is VAJAX

Variable	Coefficient	Std. Error	T-Stat.	2-Tail Sig.
C	-85998.8	22064.527	-3.8976051	0.000
SUELD1	2689.53	449.4115	5.9845608	0.000
GWH	329.7852	121.5951	2.7121586	0.009

R-squared	0.554879	Mean of dependent var	34853.89
Adjusted R-squared	0.537423	S.D. of dependent var	23785.39
S.E. of regression	16,177.16	Sum of squared resid	1.33D+10
Durbin-Watson stat	1.456968	F-statistic	31.78775
Log likelihood	-598.4126		

12. LOS SUPUESTOS DEL MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS Y MÉTODOS ALTERNATIVOS DE ESTIMACIÓN

Cuando se expuso el método de mínimos cuadrados ordinarios se establecieron una serie de supuestos que deberían cumplirse para su adecuada aplicación. Veremos a continuación que es posible contar con alternativas para estimar modelos de regresión cuando estos supuestos no se cumplen. Se indicarán estas alternativas sin pretender discutirlas a fondo.

1. El método de MCO supone que la ecuación de regresión está correctamente especificada, es decir que *existe una relación lineal.* Sin embargo, es posible aplicar este método a algunas relaciones no lineales luego de realizar transformaciones algebraicas. Como ejemplo consideremos dos casos de relación no lineal: la relación exponencial y la ecuación de la parábola.

Relación exponencial

La ecuación exponencial se define de la siguiente manera:

$$Y = AX^b \tag{34}$$

donde A y b son los coeficientes de regresión, cuyos valores se desea estimar. Tomando logaritmos a ambos lados de la ecuación exponencial, tenemos:

$$\log Y = \log (AX^b)$$
$$\log Y = \log A + b \log X \tag{35}$$

Definiendo:

$$y = \log Y$$
$$x = \log X$$
$$a = \log A$$

y sustituyendo en la ecuación (35), obtenemos una ecuación lineal en x e y:

$$y = a + bx \tag{36}$$

Ahora podremos aplicar el método de mínimos cuadrados ordinarios para estimar los coeficientes de regresión a y b de la ecuación (36). Subsecuentemente, el coeficiente original A se calcula como el antilogaritmo de a.

Relación parabólica

La ecuación de la parábola de segundo grado es la siguiente:

$$Y = a + bX + cX^2 \tag{37}$$

Definiendo la siguiente transformación: X = Z, convertimos la ecuación (37) en una ecuación de regresión lineal múltiple:

$$Y = a + bX + cZ \qquad (38)$$

Los coeficientes a, b y c pueden estimarse usando el método de MCO. En el cálculo de estos coeficientes se pierden tres grados de libertad.

2. El método de MCO supone que *x es una variable no aleatoria*. Si no se cumple este supuesto, este método ya no produce los mejores estimadores lineales no-sesgados. Será necesario usar otro método de estimación. La econometría ha desarrollado una variedad de métodos de estimación que se adecuan a los casos en que el modelo de regresión no satisface los supuestos del método de MCO. Así, si X es aleatoria, es decir si es dependiente de otra(s) variable(s), existirá una ecuación adicional que la explica. Esto origina lo que se conoce como el problema de *simultaneidad*, donde se tiene más de una ecuación. La estimación de ecuaciones simultáneas puede realizarse mediante diferentes métodos disponibles; por ejemplo, el método de mínimos cuadrados bietápicos, mínimos cuadrados trietápicos o el método de variables instrumentales. Estos métodos se discuten en detalle en textos de econometría. Muchos programas computarizados de regresión tienen la opción de usar algunos de estos métodos; el TSP ofrece la posibilidad de usar el método de mínimos cuadrados bietápicos.

3. Otro de los supuestos del método de MCO es que los *valores de Y son independientes unos de otros*. Si este supuesto no se satisface se tiene el problema de *autocorrelación* o correlación en serie. Este problema tiende a ocurrir cuando se usan series de tiempo, también llamadas series históricas o cronológicas. La autocorrelación se detecta calculando el estadístico de Durbin-Watson, y cuando esta existe el método de MCO ya no produce el mejor estimador lineal no-sesgado. En estos casos, para esti-

mar el modelo de regresión será necesario utilizar el método de mínimos cuadrados generalizados.

4. El método de MCO también supone que la desviación estándar de la distribucción de Y es constante para diferentes valores de X, característica que se denomina homoscedasticidad. Cuando este supuesto no se cumple enfrentamos un problema de heteroscedasticidad, y los estadígrafos estimados con el método de MCO no son los mejores estimadores lineales no-sesgados. La heteroscedasticidad ocurre con frecuencia cuando los datos provienen de un estudio de *corte transversal*, también llamado corte seccional. Los datos de corte transversal son aquellos que se observan en un mismo período para diferentes valores de las variables de interés. El problema de heteroscedasticidad surge muchas veces por la omisión de variables independientes relevantes que explican la variable dependiente. Por lo tanto, será necesario revisar la especificación de la ecuación de regresión e incluir alguna de estas variables independientes importantes.

5. En el análisis de regresión múltiple suponemos que las variables independientes no están altamente correlacionadas. De no cumplirse este supuesto, enfrentamos el problema de *multicolinealidad.* Como se explicó en la sección 10, una de las maneras de resolver la multicolinealidad es omitiendo una de las variables altamente colineales de la ecuación de regresión.

13. LIMITACIONES DEL ANÁLISIS DE REGRESIÓN

En esta sección se plantean algunas limitaciones inherentes al análisis de regresión. En primer lugar, señalaremos la tendencia a identificar el concepto de correlación con el de *causa-efecto* entre las variables de la ecuación de regresión, simplemente porque una variable es tratada como independiente y la otra como dependiente. En realidad, cambios en la variable independiente no necesariamente *causan* cambios en la variable dependiente. La causalidad podría ir en la otra dirección, o podría haber un

factor común que afecte a ambas variables. Las técnicas de regresión son usadas básicamente cuando una de las variables es conocida o puede ser controlada y la otra es aleatoria y no es fácilmente controlable. Así, por ejemplo, en el estudio que busca predecir el nivel de precios, variable dependiente, sobre la base de cambios en la oferta monetaria, variable independiente, es posible que la relación causa-efecto se dé en la dirección contraria si la autoridad monetaria fijase la oferta monetaria con base en el cambio de precios.

En segundo lugar, existe el riesgo de realizar *extrapolaciones exageradas* de la variable dependiente utilizando la ecuación de regresión. Este problema se presenta cuando tratamos de predecir valores de la variable dependiente sobre la base de valores observados o proyectados de las variables independientes que están fuera del rango de los datos que se utilizaron para estimar la ecuación de regresión. En el caso de la mina "La Hermosa" se estableció una relación entre los niveles de producción de mineral y las horas/hombre empleadas. Esta relación lineal sólo será válida en el rango de los datos observados. Es posible que esta relación no se mantenga para valores de la variable independiente que se encuentren fuera del rango muestral. Puede ocurrir que al incrementar el número de horas/hombre empleadas, la productividad disminuya y se presente una situación como la que se muestra en el gráfico 43. Si este fuera el caso, no podemos usar la ecuación estimada para proyectar los valores de producción de mineral tomando como base el uso de horas/hombre mayores que dicho nivel (S).

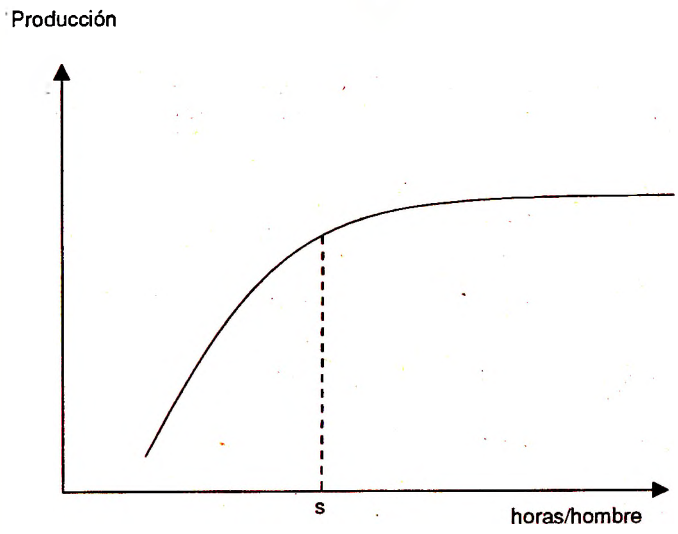
Finalmente, debemos señalar la diferencia entre los conceptos de significación versus la habilidad de estimación. Si el coeficiente de correlación $\int$ es significativamente diferente de cero, esto no necesariamente significa que podemos obtener un buen estimado de Y usando nuestra ecuación de regresión. La prueba de significación de $\int$ depende sólo de n y r $(t = r/\sqrt{(1 - R^2)/(n - k)}$, donde (k) es el número de coeficientes

de regresión). Aun con un coeficiente de correlación muestral muy pequeño, se puede lograr un ∫ significativo si n es muy grande. Por ejemplo, si r = 0.10 y n = 3,500, el estadístico t que sirve para aceptar la hipótesis de que ∫ no es significativamente diferente de cero da un valor de:

$$t = \frac{0.10}{\sqrt{(1 - 0.01)/(3{,}500 - 2)}} = \frac{0.10}{0.016} = 6.23$$

Este t calculado es mucho mayor que el t crítico, que es igual a 1.96, lo que implica que ∫ es estadísticamente significativo a pesar de que la ecuación de regresión sólo explica el 1% de las variaciones de la variable dependiente.

Gráfico 43: Curva hipotética que relaciona la producción minera con horas/hombre empleadas



Anexo: Método de mínimos cuadrados ordinarios

El método de mínimos cuadrados ordinarios (MCO) es una técnica para ajustar la línea recta "óptima" a la muestra de las observaciones de X e Y. Esto involucra minimizar la suma de los errores al cuadrado, que es una función de a y b:

$$\text{Min} \sum (Y_i - \hat{Y}_i)^2 = \text{Min } F(a, b)$$

donde Y_i se refiere a la observación i-ésima.
$\hat{Y}_i$ se refiere al valor estimado con la ecuación de regresión $Y_i = a + bX_i$

Sustituyendo el valor de $\hat{Y}_i$, tenemos:

$$\text{Min } F(a, b) = \text{Min} \sum (Y_i - a - bX_i)^2$$

La condición de primer orden para minimizar una función requiere obtener las derivadas parciales de la función con respecto a *a* y *b* e igualarlas a cero:

(i) $\dfrac{\delta F(a, b)}{\delta_a} = -2 \sum (Y_i - a - bX_i) = 0$

(ii) $\dfrac{\delta F(a, b)}{\delta_b} = -2 \sum (Y_i - a - bX_i)(X_i) = 0$

Expandiendo la ecuación (i) y reagrupando, tenemos:

$-2 \sum Y_i + 2 \sum a + 2 b \sum X_i = 0$

(iii) $\sum Y_i = na + b \sum X_i$

Expandiendo la ecuación (ii) y reagrupando, tenemos:

$$-2 \sum Y_iX_i + 2\,a \sum X_i + 2\,b \sum X_i^2 = 0$$

$$\text{(iv)} \quad \sum X_iY_i = a \sum X_i + b \sum X_i^2$$

donde n es el número de observaciones, y *a* y *b* son los estimadores de MCO.

Las segundas derivadas son positivas, condición suficiente que indica que se obtuvo un punto mínimo de F(a,b). Las ecuaciones (iii) y (iv) son conocidas como las "ecuaciones normales". Resolviendo simultáneamente estas ecuaciones, obtenemos:

$$\text{(v)}\ b = \frac{n \sum X_iY_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\text{(vi)}\ a = \frac{\sum Y_i}{n} - b\,\frac{\sum X_i}{n} = \overline{Y} - b\overline{X}$$

Ejercicios

1. Explique las siguientes afirmaciones:

a. Existe una diferencia conceptual clara entre el coeficiente de determinación y el coeficiente de correlación simple.

b. Una ecuación de estimaciones sólo es válida para el rango de valores usados para su desarrollo.

c. La multicolinealidad es un problema que sólo existe en el análisis de regresión múltiple.

d. Para determinar si un coeficiente de regresión es significativo se usa una tabla t.

e. En el análisis de regresión múltiple, al aumentar una variable al conjunto de variables explicativas siempre se obtendrá un incremento en el coeficiente de determinación múltiple.

2. Partiendo de una muestra de 200 pares de observaciones, se calcularon las siguientes cantidades:

$$\sum X = 11.34 \qquad \sum Y = 20.72 \qquad \sum X^2 = 12.16$$
$$\sum Y^2 = 84.96 \qquad \sum XY = 22.13$$

a. Estimar la línea de regresión de Y respecto a X.

b. Estimar la línea de regresión de X respecto a Y.

c. Estimar la varianza del coeficiente de regresión estimado de la regresión de Y respecto a X.

3. Las siguientes sumas fueron obtenidas a partir de dieciséis pares de observaciones de X e Y.

$$\sum X = 96 \qquad \sum Y = 64 \qquad \sum X^2 = 657$$
$$\sum Y^2 = 526 \qquad \sum XY = 492$$

a. Estimar la regresión de Y en X.

b. Contrastar la hipótesis de que la pendiente sea 1.

4. La siguiente tabla presenta los datos para una muestra aleatoria de doce parejas, sobre el número de hijos que tienen, Y_i, y el número de hijos que habían planeado tener al momento del matrimonio, X_i.

Pareja	1	2	3	4	5	6	7	8	9	10	11	12
Y_i	4	3	0	4	4	3	0	4	3	1	3	1
X_i	3	3	0	2	2	3	0	3	2	1	3	2

a. Estimar la línea de regresión de Y respecto a X.
b. Calcular la varianza del coeficiente de regresión estimado en (a).
c. Estimar el valor medio de Y correspondiente a un valor de X fijado en X = 5 y hallar un intervalo de confianza de 95% para esta media.

5. Un analista está estudiando la relación entre la distancia en kilómetros que existe entre la planta y el punto de destino (X) y el tiempo de entrega de la mercadería, en días (Y). Estableció la siguiente línea de regresión basándose en una muestra aleatoria de diez embarques:

$$\hat{Y} = 0.11 + 0.036\ X$$

Con los siguientes estadísticos:

$\bar{X} = 762$	Rango de X: [215, 1,350]
$\bar{Y} = 2.85$	Rango de Y: [1.0, 5.0]
$\sum (X_i - \bar{X})^2 = 1'297{,}860$	
$S_{yx} = EEE = 0.46$	
$S_b = 0.0004$	

a. Estime el tiempo de entrega de un embarque cuyo punto de destino se encuentra a 1,000 kilómetros de distancia.
b. ¿Se podría usar la ecuación de regresión para estimar el tiempo de entrega de un embarque cuyo punto de destino está a 2,500 kilómetros de distancia?
c. Construya un intervalo de estimación del 95% de confianza para el tiempo de entrega de un embarque que debe trasladarse a 1,000 kilómetros de la planta.
d. Determine un intervalo de confianza del 95% para la estimación del parámetro de la pendiente de la ecuación de regresión (β).
e. Pruebe la hipótesis nula de que $H_o: \beta = 0$, a un nivel de significación del 5%.

6. Los siguientes datos corresponden a siete empleados escogidos al azar y se refieren a estadísticas del año pasado. La variable X representa

el número de ausencias en días laborables, mientras que la variable Y representa la antigüedad en la compañía, expresada en años:

Y	2	0	5	6	4	9	2
X	7	8	2	3	3	3	7

a. Estime la relación lineal para estos datos.
b. Grafique los siete puntos y la línea obtenida.
c. Calcule el EEE.
d. ¿Presentan los datos suficiente evidencia de que "Y" y "X" están linealmente relacionadas? Pruebe la hipótesis de que $\beta = 0$, usando $\alpha = 0.05$.

7. La siguiente función de consumo fue estimada al estudiar una muestra de 100 unidades familiares en el área de "Villa Salvado". Y representa el gasto de consumo en intis de 1987, mientras X representa el ingreso de la unidad familiar en intis de 1987.

$$Y = 1{,}800 + 0.75\,X$$

$$S_{yx} = 4{,}500 \qquad S_x = 4{,}800 \qquad \bar{Y} = 10{,}200$$

a. ¿Cuál es el ingreso promedio de la muestra?
b. ¿Cuál es el coeficiente de correlación r?
c. ¿Qué porcentaje de la variación de Y es explicado por X?
d. ¿Cuál es el consumo promedio de todas las unidades familiares con ingresos iguales a I/.16,000?

8. Se realiza el cálculo de regresión para determinar la relación entre el precio de las acciones de la compañía "Minas Ventura" (Y) y el precio promedio del mercado de valores (X). Los resultados para 23 observaciones fueron:

$$\sum X = 5{,}414 \qquad \sum Y = 548 \qquad \sum X^2 = 1'388{,}100$$

$$\sum Y^2 = 14{,}391 \qquad \sum XY = 140{,}740$$

a. Calcule los coeficientes de regresión.
b. Calcule el error estándar de estimación.
c. Con un nivel de significación del 1%, realice un contraste de hipótesis para determinar si el parámetro de la pendiente es mayor que 0.10.
d. Construya un intervalo de confianza del 95% para el parámetro de la pendiente.

e. Construya un intervalo de predicción del 95% de confianza para el precio de la acción de "Minas Ventura" cuando el promedio del mercado de valores es 250.

9. El Gerente de Ventas de una compañía que comercializa pan integral a través de una cadena de supermercados está interesado en estudiar la relación que existe entre el precio al mayoreo de su producto y la publicidad, con las ventas alcanzadas. Para esto, registró las ventas anuales (Y) en miles de unidades, el precio unitario promedio al mayoreo (X_1) y los gastos en publicidad (X_2) en cada una de las 25 zonas durante el año pasado.

Se utilizó el programa de regresión TSP para ajustar el siguiente modelo lineal a los datos obtenidos para las 25 zonas de operación de la compañía de comercialización:

$$\hat{Y} = a + bX_1 + cX_2$$

Los resultados aparecen a continuación:

Variable	Coeficiente	Desviación estándar	Valor t
Constante	35.617		
X_1	-72.821	26.861	-2.711
X_2	3.346	0.564	5.933

R^2 = 0.6857
EEE = 4.6604

a. ¿Proporciona el modelo propuesto un buen ajuste a los datos?
b. Interprete los estimados puntuales de los dos coeficientes de regresión.
c. ¿Son los coeficientes de regresión significativamente diferentes de cero, a un nivel de significación del 5%?

10. Un estudio que intenta explicar la demanda de bebidas alcohólicas presenta la siguiente ecuación de regresión estimada tomando como base 20 observaciones anuales:

$$\hat{Y} = \underset{(0.012)}{0.014} - \underset{(0.269)}{0.354}\, X_1 + \underset{(0.00053)}{0.001}\, X_2 + \underset{(0.0034)}{0.0059}\, X_3$$

$$\bar{R}^2 = 0.68$$

donde Y Cambio anual en el consumo de alcohol por adulto.
X_1 Cambio anual en el precio real de las bebidas alcohólicas.
X_2 Cambio anual en el ingreso real disponible por persona.
X_3 Cambio anual en los gastos de propaganda en bebidas alcohólicas por adulto.

Los valores en paréntesis debajo de los coeficientes estimados son sus errores estándar.

a. Analizando el coeficiente de determinación ajustado, ¿cuán bien explican las tres variables independientes la variación de la alteración dependiente?

b. Interprete los estimados puntuales de los tres coeficientes de regresión.

c. Encuentre el intervalo de confianza del 90% para β_1.

d ¿Es la variable X_3 estadísticamente significativa para explicar variaciones en la variable dependiente, al 5% de significación?

11. A continuación se presentan las calificaciones que obtuvieron quince estudiantes que tomaron un curso de Métodos Cuantitativos después de haber seguido el curso de Matemáticas. Se desea saber si existe alguna relación entre las notas que obtuvieron los estudiantes en ambas asignaturas:

Estudiante	Matemáticas	Métodos Cuantitativos
1	80	88
2	77	70
3	50	55
4	96	95
5	40	53
6	63	75
7	60	61
8	69	78
9	95	87
10	84	80
11	89	90
12	58	67
13	71	60
14	77	79
15	72	70

Sobre la base de esta información, y utilizando la calificación obtenida en el curso de Matemáticas como la variable independiente:

a. Estime la ecuación de regresión lineal.

b. Calcule el valor esperado para la nota en el curso de Métodos Cuantitativos cuando la calificación en Matemáticas es 80.

c. Encuentre el error estándar de estimación.

d. Halle el coeficiente de determinación.

e. Pruebe la significación del coeficiente de la variable independiente, utilizando un nivel de significación de 0.05.

12. La compañía de investigación de opinión pública "Peruanísima" lleva a cabo estudios de hogares regularmente, utilizando cuestionarios que envía por correo. La compañía quiere determinar los factores que influyen en la tasa de respuesta. Se diseñó un experimento enviando treinta conjuntos de cuestionarios con diferente número de preguntas y diferentes tamaños. El modelo de regresión planteado fue:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \mu$$

donde	Y	Porcentaje de cuestionarios recibidos.
	X_1	Número de preguntas planteadas.
	X_2	Longitud del cuestionario, en términos de número de palabras.

Se estimaron los coeficientes del modelo de regresión con el programa TSP:

$$\hat{Y} = 74.3652 - \underset{(0.6349)}{1.8345}\, X_1 - \underset{(0.0091)}{0.0162}\, X_2$$

$$R^2 = 0.637$$

Los valores en paréntesis debajo de los coeficientes estimados son sus errores estándar.

a. Interprete los coeficientes estimados.

b. Calcule el valor del coeficiente de determinación.

c. Encuentre e interprete el intervalo de confianza del 99% para β_1.

d. ¿Es la variable X_2 estadísticamente significativa al 5%?

13. Una compañía inmobiliaria quiere establecer una manera adecuada para poner un precio justo a las casas que vende. Decide usar el

análisis de regresión para fijar sus precios. El modelo de regresión lineal propuesto tiene como variable dependiente el precio de la casa en miles de dólares y como variables independientes el número de dormitorios (X_1) y el área del terreno en metros cuadrados (X_2). Se utilizó una muestra aleatoria de diez casas y se estimaron los coeficientes del modelo de regresión lineal utilizando el paquete estadístico TSP. Basándose en los reportes de salida del TSP, conteste a las siguientes preguntas.

LS // Dependent Variable is Y

SMPL 1 - 10

10 Observations

Variable	Coefficient	Std. error	T-stat	2-Tail sig.
C	-5.2093023	5.3731443	-0.9695072	0.365
X_1	21.267442	1.5902260	13.373849	0.000
X_2	0.1010610	0.0089266	11.321341	0.000

R-squared	0.984993	Mean of dependent var	100.3000
Adjusted R-squared	0.980705	S.D. of dependent var	37.53532
S.E. of regression	5.213905	Sum of squared resid	190.2936
Durbin-Watson stat	1.732107	F-statistic	229.7204
Log likelihood	-28.91930		

a. Escriba la ecuación de mínimos cuadrados.
b. ¿Cuál es el valor del EEE? ¿Le indica este valor un buen ajuste?
c. ¿Cuál es el valor del coeficiente de determinación múltiple? ¿Le indica este valor un buen ajuste?
d. ¿Es la variable X_1 estadísticamente significativa al 1%?
e. Tomando como base sus respuestas (b) - (d), ¿debe la compañía adoptar el modelo propuesto?
f. Suponga que el modelo es adoptado. ¿Cuál es el estimado puntual del precio de una casa de dos dormitorios en un terreno de 400 metros cuadrados?
g. Construya un intervalo de confianza del 98% para el precio de la casa descrita en (f).

14. Una corporación multinacional que produce *chips* de silicona para computadoras está soportando serios problemas de robos por los mismos trabajadores de la empresa. Dado que los *chips* son muy pequeños no existe una manera definitiva de evitar los robos. La compañía

propone contratar guardias de seguridad, instalar monitores y utilizar perros policía en un intento de disminuir los cuantiosos robos. La compañía seleccionó diez de sus plantas de producción y en cada una implementó una combinación de guardias (X_1), monitores (X_2) y perros (X_3). Finalmente, se registró el número de miles de *chips* que fueron robados en un mes (Y).

El Director de Investigación de la compañía utilizó el programa estadístico TSP para estimar los coeficientes del modelo propuesto, obteniendo el siguiente reporte:

LS // Dependent Variable is Y

SMPL 1 - 10

10 Observations

Variable	Coefficient	Std. error	T-stat	2-Tail sig.
C	25.599719	9.5721934	2.6743838	0.037
X_1	-0.1206355	0.7338819	-0.1643799	0.875
X_2	-1.1152497	0.4928550	-2.2628355	0.064
X_3	0.2536887	0.3838758	0.6608613	0.533

R-squared	0.498910	Mean of dependent var	14.20000
Adjusted R-squared	0.248365	S.D. of dependent var	6.390966
S.E. of regression	5.540769	Sum of squared resid	184.2008
Durbin-Watson stat	1.857290	F-statistic	1.991297
Log likelihood	-28.75659		

a. Escriba la ecuación de mínimos cuadrados.
b. ¿Cuál es el valor del coeficiente de determinación múltiple? ¿Le indica este valor un buen ajuste?
c. ¿Son las variables independientes estadísticamente significativas al 5%?
d. Con base en las respuestas (b) y (c), ¿debe la compañía adoptar el modelo de regresión propuesto?
e. ¿Qué otra información le gustaría ver en las salidas del programa TSP antes de contestar la pregunta (d)?
f. Suponga que el modelo de regresión propuesto es adoptado. ¿Cuál es el estimado puntual del número de miles de *chips* robados en una planta dotada con 20 guardias, 50 monitores y 20 perros policía?

15. El Gerente de Ventas de una compañía farmacéutica está preocupado por un aparente menor rendimiento de sus agentes más expe-

rimentados. Ha observado que mientras más años de experiencia tienen, sus ventas no sólo se estabilizan sino que en algunos casos decrecen. Para estudiar este problema, el Gerente ha seleccionado diez territorios de venta y ha registrado las ventas de los tres últimos meses, en miles de dólares, así como la experiencia, en años, de los agentes responsables de cada uno de los territorios.

Ventas (En miles de $)	Experiencia (Años)	Ventas (En miles de $)	Experiencia (Años)
36.7	2.0	41.2	4.5
22.9	1.5	18.5	1.0
30.5	4.5	43.4	3.0
9.2	0.8	25.5	2.3
38.4	3.5	28.4	5.5

a. Grafique la relación entre las ventas y la experiencia del agente.

Sobre la base del diagrama de dispersión se ajustó la relación entre ventas (Y) y experiencia del agente (X) usando un modelo polinomial de segundo orden. Se utilizó el programa estadístico TSP y los resultados fueron:

Variable	Coeficiente	Desviación estándar	Valor t
Constante	-6.1730		
X_1	25.332	5.439	4.657
X_2	-3.4990	0.872	-4.014

R^2 = 0.8029
EEE = 5.4526

b. ¿Está de acuerdo con el modelo propuesto? ¿Por qué?
c. ¿Qué proporción de la variación en las ventas es explicada por el modelo?
d. ¿Son los coeficientes de regresión significativamente diferentes de cero, con un nivel de significación del 5%?

VII. Predicción

1. Componentes de una serie de tiempo. 2. Predicciones con series de tiempo. 3. Predicciones de series de tiempo usando la técnica de descomposición. 4. Predicciones con métodos causales. 5. Predicciones cualitativas.

En el proceso de toma de decisiones, tanto a nivel individual como de empresas e instituciones públicas, existen factores o variables que afectan el resultado de la decisión pero que están fuera del control del decisor. La predicción del comportamiento futuro de estas variables se torna una actividad crucial para la formulación de estrategias apropiadas.

Así, por ejemplo, la decisión de un individuo de seguir o no estudios de postgrado se basa implícita o explícitamente en sus predicciones sobre sus oportunidades futuras en el mercado laboral. Asimismo, una empresa tiene la necesidad de predecir sus ventas y costos futuros para poder decidir sus niveles de producción, inventarios, compra de materias primas, contratación de personal, etcétera. De igual manera, las instituciones públicas basan sus políticas en estimaciones del futuro. La política monetaria del Banco Central de Reserva dependerá de las predicciones relacionadas con el crecimiento de la economía y los futuros niveles de inflación.

Las *predicciones* pueden clasificarse como predicciones de *corto, mediano* y *largo plazo,* dependiendo de cuán lejos en el futuro se desee predecir. Esta clasificación es relativa, y depende del decisor y del tipo de problema bajo análisis. Así, mientras que para una determinada empresa una predicción de corto plazo implica predecir el próximo año, para un vendedor dedicado al comercio ambulatorio esta significará predecir el día siguiente. Por convención, a nivel empresarial y de gobierno, corto plazo implica predicciones de una semana hasta un año; predicciones de mediano plazo son las de cinco años; y predicciones de largo plazo son las de diez a quince años. En general, dado que la distinción entre corto y largo plazo varía, en cada situación de predicción debemos especificar claramente lo que consideramos como corto, mediano y largo plazo.

Existen muchos *métodos de predicción,* los cuales pueden ser clasificados en tres grandes categorías: métodos de series de tiempo, métodos causales y métodos cualitativos. Los *métodos de series de tiempo* se basan en el análisis de los datos históricos de la variable que estamos tratando de proyectar. En este análisis, el tiempo es la variable independiente. Para que estas predicciones sean adecuadas se debe disponer de un número suficiente de observaciones históricas, y debe tratarse de una variable con un patrón de comportamiento relativamente estable, dado que estos métodos suponen que tales patrones permanecerán vigentes en el futuro.

Los *métodos causales* relacionan la variable que intentamos predecir con otras variables diferentes al tiempo. Un ejemplo de estos métodos es el análisis de regresión presentado en el capítulo anterior.

Los *métodos cualitativos* de predicción se basan primordialmente en información cualitativa, como es la opinión de expertos en el tema de interés. La ventaja de estos métodos es que pueden aplicarse en situaciones en las que no se dispone de datos históricos.

1. COMPONENTES DE UNA SERIE DE TIEMPO

Una *serie de tiempo,* también llamada serie cronológica, se define como un conjunto de valores de una variable específica, registrados en períodos sucesivos. El patrón de comportamiento de los datos de una serie de tiempo está determinado por la combinación de varios componentes. En algunas series de tiempo se puede identificar hasta cuatro diferentes componentes: tendencia secular, variaciones cíclicas, variaciones estacionales y fluctuaciones irregulares. El análisis visual de una serie de tiempo mediante su representación en un diagrama de dispersión no es sencillo, debido a que los componentes básicos aparecen entremezclados. El objetivo de esta sección es definir claramente cada uno de estos componentes, para luego intentar descomponer la serie en los patrones básicos que la afectan.

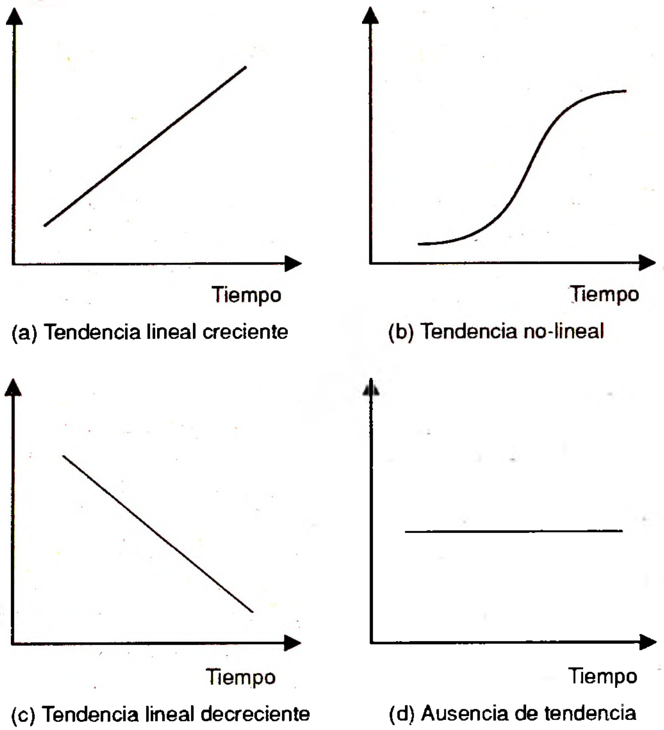
A. Tendencia secular

Las observaciones de una serie de tiempo se pueden haber tomado cada hora, día, semana, mes, año o a cualquier otro intervalo regular de tiempo. A pesar de que estas observaciones generalmente muestran fluctuaciones irregulares, es posible que presenten cambios graduales o movimientos a valores mayores o menores en un período largo. Este movimiento gradual es conocido como la *tendencia secular* de la serie de tiempo. Esta tendencia se debe usualmente a factores de largo plazo como los cambios en la población, cambios en las características demográficas de la población, cambios en las preferencias del consumidor y cambios tecnológicos.

El gráfico 44 muestra diferentes tipos de tendencia secular. En la parte (a) se presenta una tendencia lineal creciente. Esta figura bien podría representar la demanda de baterías ante un crecimiento del parque automotor a través del tiempo. En (b) se muestra una tendencia no lineal. Esta curva describe una serie de tiempo que tiene poco crecimiento inicialmente, seguida de un período de

rápido crecimiento, y luego una etapa de estabilización. Esta puede ser una buena descripción de las ventas de un nuevo producto desde su introducción, seguida de un rápido crecimiento para terminar con un período de saturación del mercado. La tendencia lineal decreciente en (c) es una descripción adecuada para series de tiempo que muestran un decrecimiento continuo. La línea horizontal en (d) se usa para series de tiempo que no muestran ni un crecimiento, ni un decrecimiento consistente a través del tiempo; en este caso se dice que no existe una tendencia secular.

Gráfico 44: Algunos patrones de tendencia secular



B. El componente cíclico

Si bien una serie de tiempo muestra una variación gradual o tendencia secular en períodos largos, no se puede esperar que todos los valores de la serie de tiempo estén exactamente en la línea de tendencia. A pesar de que las observaciones sean tomadas con un intervalo de un año, las series de tiempo muestran a menudo secuencias alternas de puntos por debajo y encima de la línea de tendencia. Cualquier patrón regular de secuencias de puntos por debajo y encima de la línea de tendencia es atribuible al *componente cíclico* de la serie de tiempo. Se piensa que estos movimientos representan ciclos de varios años de duración, y que para estudiarlos se necesitan observaciones de por lo menos quince años.

C. Variaciones estacionales

Mientras que los componentes de tendencia y cíclicos de una serie de tiempo se identifican analizando movimientos de datos históricos en varios años, muchas series de tiempo muestran un patrón regular de variación en el transcurso de un año. Por ejemplo, los trajes de baño se venden muy poco en los meses de otoño e invierno, mientras que durante los meses de primavera y verano las ventas alcanzan su punto más alto. Los fabricantes de chocolates esperan un patrón de ventas exactamente opuesto. Entonces, no debe sorprender que el componente de las series de tiempo que representan las variaciones en los datos debido a las influencias estacionales se llame *componente estacional*. Pero este componente no sólo es atribuible a variaciones del clima, sino que también puede deberse a costumbres o características especiales del producto. Un caso típico es la venta de panetones en Lima, que alcanzan sus niveles más altos en épocas de Fiestas Patrias (julio) y Navidad (diciembre).

Aunque comúnmente se piensa que las variaciones estacionales de una serie de tiempo ocurren durante un año, el componente estacional puede también representar cualquier patrón que se repite en un período menor a un año. Por ejemplo, el volumen de ventas de alimentos en un supermercado alcanza sus niveles más altos durante los fines de semana, y este es un patrón que se repite cada semana.

D. Fluctuaciones irregulares

Supongamos que tenemos una serie de tiempo donde hemos identificado la tendencia secular, las variaciones cíclicas y las variaciones estacionales. Existirá todavía un factor residual que expresa la desviación entre el valor observado y el valor explicado por los tres componentes antes mencionados. Estos residuos representan las *fluctuaciones irregulares* que son causadas por factores enteramente fortuitos, de corto plazo, y que no son recurrentes. Este componente es impredecible, dado que representa fluctuaciones aleatorias.

Debemos señalar que a medida que se acortan los intervalos de tiempo de observación de los datos, mayores son las fluctuaciones irregulares que afectan a la serie de tiempo. Así, si las ventas se observan diariamente, es de esperar que los datos muestren mayores efectos del componente irregular. En cambio, si las ventas diarias se agregan para formar observaciones semanales, las fluctuaciones irregulares tenderán a cancelarse unas con otras, y serán menos importantes. De igual manera, si comparamos series de tiempo de variables desagregadas versus variables agregadas, veremos que las primeras muestran una mayor influencia de las variaciones aleatorias. Por lo general, las series estadísticas correspondientes a una empresa individual son más propensas a presentar irregularidades que las referentes a toda la industria. Cuando se combinan los datos de un gran número de empresas individuales es muy probable que se cancelen las irregularidades individuales.

2. PREDICCIONES CON SERIES DE TIEMPO

El propósito fundamental del estudio de las series de tiempo es el análisis de los datos históricos de una variable en un período determinado, para poder predecir valores futuros de esta variable. Existe una serie de métodos que aplican un conjunto de técnicas replicables para lograr dicho estudio. Parte del arte de la predicción es seleccionar la técnica, o conjunto de técnicas, más apropiadas y efectivas. Usualmente, es necesario realizar un análisis extenso de las series de tiempo a fin de obtener la información máxima, puesto que estas pueden estar influenciadas por todos los componentes mencionados en la sección anterior.

En el análisis de series cronológicas, el tiempo es la variable independiente y se representa en el eje de las abscisas (X) en los gráficos de dispersión.

A. Técnicas de predicción de corto plazo

Aquí discutimos las técnicas de predicción apropiadas para series de tiempo que no muestran tendencias significativas, ni fluctuaciones cíclicas, ni variaciones estacionales. En tales situaciones, el objetivo del método de predicción es "suavizar" las fluctuaciones irregulares de la serie de tiempo. Estas técnicas intentan predecir el valor que tomará la variable en el próximo período en el futuro.

a. Predicción histórica

Se predice el próximo período con el valor del último período histórico. Es decir:

$$\hat{Y}_{t+1} = Y_t \qquad (1)$$

donde $\hat{Y}_{t+1}$ es el valor estimado para el período t+1.
Y_t es el valor observado en el período t, siendo t el último período histórico.

b. Media aritmética

Este método consiste en calcular la media aritmética de todos los valores de la serie de tiempo y utilizarla como la predicción para el próximo período.

$$\hat{Y}_{t+1} = \frac{\sum Y_{t-1}}{n} \qquad (2)$$

donde $\hat{Y}_{t+1}$ es el valor estimado para el período t+1.
n es el número total de observaciones de la variable Y.
Y_{t-1} es el valor observado de Y en el período t-1.

En algunos casos, cuando las series de tiempo consisten de observaciones mensuales o trimestrales y no presentan una tendencia secular pero sí algún tipo de variación estacional, las predicciones del próximo mes o trimestre podrán realizarse utilizando la media aritmética de los mismos meses o trimestres de los años históricos.

Para ilustrar esta técnica consideremos las ventas de tortas de la "Pastelería Buena Ventura" en las doce últimas semanas, que se presentan en el cuadro 18. Nuestro objetivo es utilizar estos datos para predecir las ventas de la decimotercera semana.

Calculemos la media aritmética de los doce valores históricos:

$$\hat{Y}_{13} = Y = \frac{\sum Y_i}{12} = 71.25 \approx 71$$

Nuestra predicción para la decimotercera semana será 71 tortas.

CUADRO 18: "PASTELERÍA BUENA VENTURA": VENTAS DE TORTAS
(Últimas 12 semanas)

Semana	Ventas de tortas (Unidades)
1	69
2	73
3	71
4	75
5	70
6	68
7	72
8	70
9	74
10	72
11	67
12	74

c. Media móvil

El método de la media móvil consiste en calcular la media aritmética de los k datos más recientes en la serie de tiempo. Esta medida es utilizada para proyectar el próximo período.

$$\text{Media móvil} = \frac{\sum \text{(de los k datos más recientes)}}{k} \qquad (3)$$

Se le llama media móvil porque a medida que se obtiene una nueva observación para la serie de tiempo, esta reemplaza a la observación más antigua en la ecuación (3), calculándose una nueva media aritmética. Entonces, la media

cambiará o se "moverá" a medida que se obtengan nuevas observaciones.

Para ilustrar el método de la media móvil, consideremos el ejemplo de las ventas de tortas de la "Pastelería Buena Ventura" presentadas en el cuadro 18. Debemos elegir primero el número de datos (k) que serán incluidos en el cálculo de las medias. Como ejemplo, calculemos la predicción de la decimotercera semana basándonos en una media móvil de tres semanas:

$$\hat{Y}_{13} = \frac{Y_{12} + Y_{11} + Y_{10}}{3} = 71$$

También podríamos utilizar una media móvil de cuatro semanas, y obtener:

$$\hat{Y}_{13} = \frac{Y_{12} + Y_{11} + Y_{10} + Y_{9}}{4} = 71.75 \approx 72$$

Como vemos en este ejemplo, al utilizar medias móviles de diferente longitud obtenemos predicciones diferentes. A través de un método de tanteos es posible escoger la longitud apropiada que minimice los errores de predicción que se definen como las diferencias entre los valores observados y los estimados.

Para escoger la longitud óptima se aplica el método de los tanteos a los datos de la serie histórica. El cuadro 19 ilustra los cálculos de la media móvil de tres semanas para nuestro ejemplo de la venta de tortas. La predicción de las ventas de la cuarta semana se hace sobre la base del promedio de ventas de las tres semanas anteriores, y así sucesivamente. Luego se calculan los errores de predicción de cada semana, que se presentan en la cuarta columna de dicho cuadro.

Una forma de medir la precisión global que se logra utilizando medias móviles para efectuar predicciones sería la suma de

los errores de predicción a lo largo del período histórico. El problema con esta medida es que, puesto que los errores son aleatorios, algunos serán positivos y otros negativos, y su suma tenderá a cero, independientemente del tamaño de los errores individuales. De hecho, si observamos el cuadro 19 la suma de los errores de predicción es cero. Esta dificultad puede evitarse calculando el cuadrado de cada error de predicción.

CUADRO 19: CÁLCULOS DE LA MEDIA MÓVIL DE TRES SEMANAS

Semana	Valores de la serie de tiempo	Predicción con media móvil	Error de predicción	Errores al cuadrado
1	69			
2	73			
3	71			
4	75	71	4	16
5	70	73	-3	9
6	68	72	-4	16
7	72	71	1	1
8	70	70	0	0
9	74	70	4	16
10	72	72	0	0
11	67	72	-5	25
12	74	71	3	9
		Totales	0	92

Para poder comparar la precisión de las predicciones utilizando diferentes longitudes de medias móviles calculamos el error cuadrado promedio (ECP) para los respectivos valores de k en el período histórico:

$$ECP(k) = \frac{\sum (Y_i - \hat{Y}_i)^2}{t - k} \qquad (4)$$

donde t es el número de observaciones históricas.
k es la longitud de la media móvil.
$(Y_i - \hat{Y}_i)$ es el error de predicción para el período i, $i = k+1, k+2, \ldots, t$.

Utilizando la ecuación (4) podemos calcular el ECP para varios valores de k, y escoger la longitud de la media móvil que minimice esta medida. Del cuadro 19 calculamos:

$$ECP(3) = \frac{92}{9} = 10.22$$

Al calcular y comparar los ECP(k) para diferentes valores de k, encontramos que la media móvil de longitud 6 es la que minimiza el error cuadrado promedio. Entonces podemos predecir las ventas de la decimotercera semana como:

$$\hat{Y}_{13} = \frac{Y_{12} + Y_{11} + Y_{10} + Y_9 + Y_8 + Y_7}{6} = 71.5$$

d. Suavización exponencial

Esta técnica utiliza todos los datos disponibles de la serie histórica para calcular el promedio, base de la predicción, pero dando mayor ponderación a los datos más recientes. Estas ponderaciones dependen de un *factor de ponderación o suavización*, denotado por α. El modelo básico de suavización exponencial es el siguiente:

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \ldots + \alpha(1-\alpha)^{t-1} Y_1 \qquad (5)$$

donde $\hat{Y}_{t+1}$ es la predicción hecha en el período t para el período t+1.

Y_t es el dato del período t.

Y_{t-1} es el dato del período t-1.

α es el factor de suavización ($0 < \alpha < 1$).

t es el número de períodos de observación de la serie de tiempo.

Nótese que los coeficientes –α, $\alpha(1-\alpha)$, $\alpha(1-\alpha)^2$...– disminuyen exponencialmente a medida que se retrocede en el tiempo. Además, cuanto mayor sea el valor de α, mayor será el peso que se asigne a los datos más recientes. El problema consiste en escoger el α adecuado para calcular la ecuación (5), ya que utilizando diferentes valores de α se obtendrán diferentes predicciones. Para determinar el valor deseado de α se puede utilizar el mismo criterio considerado antes para determinar el número de períodos a incluir en el cálculo de la media móvil. Es decir, escogeremos el valor de α que minimice el ECP.

Para ilustrar la importancia del factor de suavización exponencial reescribamos la ecuación (5) para el período t:

$$\hat{Y}_t = \alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + \alpha(1-\alpha)^2 Y_{t-3} + \ldots \qquad (6)$$

Reemplazando (6) en (5), obtenemos:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\,[\alpha Y_{t-1} + \alpha(1-\alpha)Y_{t-2} + \ldots]$$

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t \qquad (7)$$

Con esta última ecuación podemos proyectar el período t+1, usando el valor observado y el valor proyectado para el período t. La ecuación (7) simplifica los cálculos del método de suavización exponencial, dado que sólo requiere del último dato histórico y de la predicción hecha para ese mismo período.

Reescribiendo la ecuación (7) se puede lograr un mejor entendimiento del factor de suavización:

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\hat{Y}_t = \alpha Y_t + \hat{Y}_t - \alpha\hat{Y}_t$$

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t) \qquad (8)$$

Entonces, la predicción $\hat{Y}_{t+1}$ es igual a la predicción previa, $\hat{Y}_t$, más un ajuste que es α veces el error de predicción más reciente. Si la serie de tiempo es volátil y contiene una variación aleatoria importante, es preferible usar un valor pequeño de α para no sobredimensionar el error de predicción, que se debe precisamente a las variaciones aleatorias. Por otro lado, si la serie de tiempo bajo análisis es relativamente estable, con pocas variaciones aleatorias, usar un α mayor tiene la ventaja de ajustar rápidamente las predicciones reaccionando a las condiciones cambiantes.

Para ilustrar el modelo de suavización exponencial consideremos la serie de tiempo de las ventas de tortas del cuadro 18. Como dijimos antes, el criterio para elegir el valor numérico de α es la minimización del ECP. El cuadro 20 presenta un resumen de los cálculos del ECP para la predicción de ventas usando la técnica de suavización exponencial para α igual a 0.2 y 0.3. Al no tener valores anteriores para calcular la predicción en el primer período, hacemos $\hat{Y}_1 = Y_1$.

Del cuadro 20 se desprende que el ECP con $\alpha = 0.2$ es 8.93, mientras que el ECP con $\alpha = 0.3$ es 9.35. Es decir que, para esta serie de tiempo, se obtiene una mejor predicción usando un $\alpha = 0.2$. Siguiendo este proceso de tanteos con otros valores de α, se podrá encontrar el mejor factor de suavización para la serie histórica. Suponiendo que este α será también el mejor valor para el futuro, podemos efectuar la predicción de ventas para la decimotercera semana.

CUADRO 20: CÁLCULO DEL ERROR AL CUADRADO PROMEDIO (ECP) PARA PREDECIR LAS VENTAS DE TORTAS, USANDO DIFERENTES CONSTANTES DE SUAVIZACIÓN

Semana t	Valor serie de tiempo $\hat{Y}_t$	Predicciones con $\alpha = 0.2$			Predicciones con $\alpha = 0.3$		
		Predicción $\hat{Y}_t$	Error de predicción $Y_t - \hat{Y}_t$	Error al cuadrado $(Y_t-\hat{Y}_t)^2$	Predicción $\hat{Y}_t$	Error de predicción $Y_t - \hat{Y}_t$	Error al cuadrado $(Y_t-\hat{Y}_t)^2$
1	69	69.00	0.00	0.00	69.00	0.00	0.00
2	73	69.00	4.00	16.00	69.00	4.00	16.00
3	71	69.80	1.20	1.44	70.20	0.80	0.64
4	75	70.04	4.96	24.60	70.44	4.56	20.79
5	70	71.03	-1.03	1.06	71.81	-1.81	3.28
6	68	70.83	-2.83	8.01	71.27	-3.27	10.69
7	72	70.26	1.74	3.03	70.29	1.71	2.92
8	70	70.61	-0.61	0.37	70.80	-0.80	0.64
9	74	70.49	3.51	12.32	70.56	3.44	11.83
10	72	71.19	0.81	0.66	71.59	0.41	0.17
11	67	71.35	-4.35	18.92	71.71	-4.71	22.18
12	74	70.48	3.52	12.39	70.30	3.70	13.69
			Total	98.80		Total	102.84
		ECP = 98.80/11 = 8.98			ECP = 102.84/11 = 9.35		

B. Técnicas de predicción de largo plazo

En esta subsección presentaremos dos técnicas para predecir más de un período en el futuro. En primer lugar, describiremos cómo predecir los valores de series de tiempo que muestran una tendencia lineal. En segundo término, presentaremos la técnica de descomposición que permite predecir series de tiempo que además de una tendencia marcada contienen variaciones estacionales y fluctuaciones cíclicas.

a. Predicciones de tendencia lineal

Esta técnica supone que la serie de tiempo muestra únicamente una tendencia secular y no variaciones cíclicas ni estacionales, y es especialmente útil cuando se analizan series de tiempo de observaciones anuales. Consiste en "ajustar" una línea de tendencia a las observaciones pasadas, y luego proyectar esa línea para estimar valores futuros.

Para ilustrar esta técnica consideremos las ventas de aspiradoras de una marca específica en los últimos diez años, que se muestran en el cuadro 21.

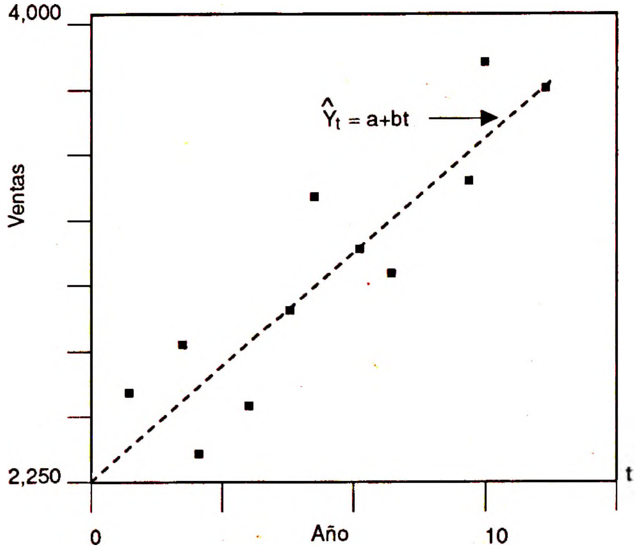
CUADRO 21: VENTAS DE ASPIRADORAS

Año t	Ventas Y_t
1	2,592
2	2,748
3	2,360
4	2,576
5	2,963
6	3,395
7	3,200
8	3,082
9	3,396
10	3,781

Los datos históricos de esta serie muestran que las ventas de aspiradoras no han aumentado en forma continua en el período considerado. En algunos años se verifican disminuciones con respecto al año anterior. Sin embargo, analizando los diez años en conjunto, se observa una tendencia creciente. El gráfico 45

ilustra este comportamiento. En el eje de las abscisas se representa el tiempo, y en el de las ordenadas el volumen de ventas.

Gráfico 45: Venta de aspiradoras



La línea punteada del gráfico 45 describe en forma apropiada las ventas de aspiradoras en el largo plazo. La función lineal que mejor se aproxime a la tendencia podrá estimarse usando la técnica de mínimos cuadrados discutida en el capítulo anterior. La variable independiente es el tiempo, y la variable dependiente es el volumen de ventas de aspiradoras. Así,

$$\hat{Y}_t = a + bt \tag{9}$$

donde $\hat{Y}_t$ es el valor de tendencia de las ventas en el período t.
a es el intercepto de la línea de tendencia.
b es la pendiente de la línea de tendencia.
t es el tiempo (t = 1, 2, 3, ... 10).

Realizando los cálculos necesarios –ver ecuaciones (5) y (6) del capítulo anterior– se obtiene la función lineal estimada de la tendencia de las ventas de aspiradoras:

$$\hat{Y}_t = 2{,}304.27 + 128t \qquad (10)$$

La pendiente de la línea de tendencia de 128 indica que en los últimos diez años las ventas han experimentado un crecimiento promedio anual de 128 aspiradoras. Si aceptamos que la tendencia de las ventas de los últimos diez años continuará en el futuro, entonces se puede usar la ecuación (10) para predecir el componente de tendencia de la serie de tiempo. Por ejemplo, si deseamos predecir la tendencia de las ventas para el próximo año, sustituimos t = 11 en la ecuación (10):

$$\hat{Y}_t = 2{,}304.27 + 128(11) = 3{,}712.27$$

El uso de una función lineal para modelar la tendencia de series de tiempo es muy común. Sin embargo, algunas series de tiempo muestran tendencias no lineales, cuyas funciones pueden ser estimadas de manera similar (ver la sección 12 del capítulo anterior).

b. Técnica de descomposición

Como se mencionó anteriormente, las series de tiempo están influenciadas por un conjunto de patrones que al aparecer entremezclados hacen difícil su análisis. La *técnica de descomposi-*

ción intenta aislar los patrones básicos de la serie de tiempo a fin de encontrar las variaciones regulares en el pasado y proyectarlas al futuro. Esta técnica será presentada en detalle en la próxima sección.

3. PREDICCIONES DE SERIES DE TIEMPO USANDO LA TÉCNICA DE DESCOMPOSICIÓN

La *técnica de descomposición* supone que la serie de tiempo bajo análisis contiene cuatro componentes: tendencia (T), estacional (E), cíclico (C) e irregular (I). También suponemos que estos componentes afectan la serie de tiempo en forma multiplicativa. Es decir, si los cuatro componentes pueden ser identificados y cuantificados, los valores de la serie de tiempo se obtendrán multiplicando estas cuatro medidas. Matemáticamente:

$$Y_t = T_t * C_t * E_t * I_t \qquad (11)$$

En este modelo, la tendencia T es medida en las unidades de la variable analizada, Y. Por otro lado, los componentes C_t e I_t son medidos en términos relativos, con valores mayores que 1.00, indicando que el efecto cíclico está por encima de la tendencia, que el efecto estacional está por encima del nivel normal o promedio, o el efecto irregular está por encima de la combinación de los componentes de tendencia, cíclico y estacional. Valores por debajo de 1.00 para C_t, E_t e I_t indican niveles por debajo de los promedios del respectivo componente.

Supóngase que con base en datos históricos de una serie de tiempo se ha logrado hacer una predicción de tendencia de 48 unidades para un período futuro (P), y que se han calculado valores de 0.90, 1.15 y 0.98 para los componentes C_P, S_P e I_P respectivamente. El valor de predicción para dicho período, utilizando el modelo multiplicativo, será:

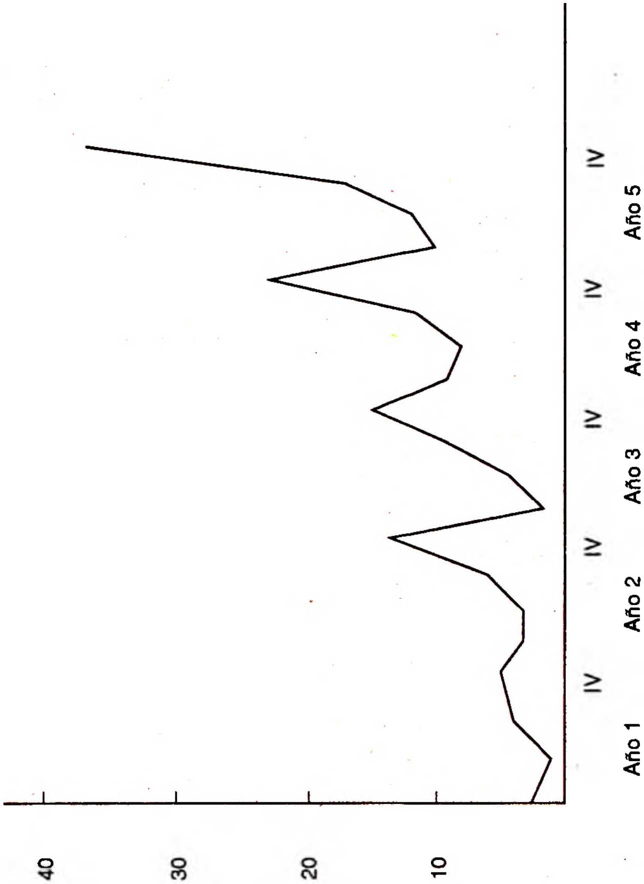
$$Y_P = 48 * 0.90 * 1.15 * 0.98 = 49$$

Ilustraremos el uso de la técnica de descomposición utilizando la serie de tiempo de las ventas trimestrales de órganos electrónicos de la compañía "Sonido Divino S.A." en los últimos cinco años. Esta información se presenta en el cuadro 22, y los datos se ilustran en el gráfico 46.

CUADRO 22: "SONIDO DIVINO S.A."
VENTAS TRIMESTRALES DE
ÓRGANOS ELECTRÓNICOS

Año / Trimestre		Ventas (Unidades)
Año 1	I	2
	II	1
	III	4
	IV	5
Año 2	I	4
	II	4
	III	6
	IV	14
Año 3	I	3
	II	5
	III	10
	IV	16
Año 4	I	9
	II	7
	III	12
	IV	22
Año 5	I	10
	II	13
	III	18
	IV	35

Gráfico 46: Serie de tiempo de la venta de órganos electrónicos



A. Estimación de los factores de estacionalidad

Al observar el gráfico 46 notamos un cierto patrón estacional en las ventas de órganos electrónicos, con ventas bajas en el primer y segundo trimestres, seguidos de ventas más altas en el tercero y cuarto trimestres. Para identificar las influencias estacionales de cada trimestre usamos un procedimiento que comienza por calcular los promedios móviles de los datos históricos. El cálculo de estos promedios elimina las influencias estacionales e irregulares (E_tI_t), y permite medir el efecto combinado de la tendencia y el componente cíclico (T_tC_t).

Al calcular promedios móviles de cuatro trimestres estamos incluyendo datos de un año en cada uno de estos promedios. Es decir, los promedios móviles siempre contienen una observación de cada uno de los cuatro trimestres del año, logrando promediar las fluctuaciones estacionales. El promedio móvil de los primeros cuatro trimestres de la serie histórica de las ventas de órganos electrónicos es:

$$\text{Primer promedio móvil} = \frac{2+1+4+5}{4} = 3.00$$

Para calcular el segundo promedio móvil añadimos las ventas del primer trimestre del año 2 excluyendo las ventas del primer trimestre del año 1:

$$\text{Segundo promedio móvil} = \frac{1+4+5+4}{4} = 3.50$$

De igual manera se calcula el resto de promedios móviles para toda la serie histórica.

Un problema con estos promedios es tratar de identificar a qué trimestre asignar estos valores. Explícitamente, tomemos el primer promedio móvil, cuyo valor es de 3.00 y representa un volumen de ventas trimestrales promedio del año 1. Es razona-

ble pensar que este promedio móvil debe asociarse al trimestre central del primer año. Es obvio que al incluir cuatro trimestres en el cálculo del promedio móvil, tal trimestre central no exista. El valor 3.00 corresponde a un período que incluye la última mitad del segundo trimestre y la primera mitad del tercer trimestre del primer año. De manera similar, el siguiente promedio móvil de 3.5 correspondería a la última mitad del tercer trimestre y la primera mitad del cuarto trimestre del año 1. Y así sucesivamente con los siguientes promedios móviles, como se muestra en la columna 3 del cuadro 23.

CUADRO 23: CÁLCULO DE LOS PROMEDIOS MÓVILES DE VENTA DE ÓRGANOS ELECTRÓNICOS

Año/Trimestre		Ventas (TCEI)	Promedios móviles de cuatro trimestres	Promedios móviles centrados (TC)	Influencias Estac.-Irreg. $EI = \frac{TCEI}{TC}$
1	I	2			
	II	1			
			3.00		
	III	4		3.25	1.23
			3.50		
	IV	5		3.88	1.29
			4.25		
2	I	4		4.50	0.89
			4.75		
	II	4		5.88	0.68
			7.00		
	III	6		6.88	0.87
			6.75		
	IV	14		6.88	2.03
			7.00		
3	I	3		7.50	0.40
			8.00		
	II	5		8.25	0.61
			8.50		

(sigue)

(viene de la pág. anterior)

	III	10		9.25	1.08
			10.00		
	IV	16		10.25	1.56
			10.50		
4	I	9		10.75	0.84
			11.00		
	II	7		11.75	0.60
			12.50		
	III	12		12.63	0.95
			12.75		
	IV	22		13.50	1.63
			14.25		
5	I	10		15.00	0.67
			15.75		
	II	13		17.38	0.75
			19.00		
	III	18			
	IV	35			

Esto quiere decir que los valores de los promedios móviles no corresponden directamente a los trimestres originales de la serie de tiempo. Podemos resolver este problema ubicando los puntos medios entre promedios móviles sucesivos. Por ejemplo, dado que 3.0 corresponde a la primera mitad del tercer trimestre, y 3.5 a la última mitad del mismo trimestre, usaremos el promedio de estos valores –(3.0+3.5)/2 = 3.25– como el valor del promedio móvil del tercer trimestre. De manera análoga, calculamos el promedio móvil para el cuarto trimestre –(3.5 + 4.25)/2 = 3.88. Y así sucesivamente para los otros trimestres. Estos valores se muestran en la columna 4 del cuadro 23, y miden la influencia combinada de la tendencia y la variación cíclica (T_tC_t) sobre los datos históricos.

Debemos señalar que si los datos de la serie de tiempo fueran mensuales, entonces los promedios móviles serían de doce meses y también habría que hacer los ajustes necesarios para centrar los promedios. En cambio, si el número de períodos

utilizados en el cálculo de los promedios móviles es un número impar, el punto "central" corresponderá a uno de los puntos originales de la serie de tiempo y no será necesario ningún ajuste. Un ejemplo de este último caso sería el estudio del comportamiento estacional de las ventas de un supermercado en los siete días de la semana.

Mientras que en la columna 2 del cuadro 23 tenemos los valores de los datos observados (Y_t) que son el resultado de la combinación de todas las influencias, en la columna 4 tenemos los promedios móviles centrados que representan el componente combinado de tendencia y cíclico (T_tC_t). Entonces, dividiendo cada observación de la serie de tiempo, Y_t, por el correspondiente promedio móvil, T_tC_t, podemos identificar el efecto estacional-irregular:

$$E_tI_t = \frac{Y_t}{T_tC_t}$$

La columna 5 del cuadro 23 muestra los valores del efecto combinado, E_tI_t. Si sólo existiera la fluctuación estacional esperaríamos que la influencia de esta en un trimestre determinado fuera la misma en cada uno de los años del período histórico. Sin embargo, dado que estos valores incluyen también la influencia irregular, vemos que varían de año a año. Por ejemplo, en el cuadro 23 se observa que los valores de E_tI_t para el cuarto trimestre varían: en el primer año es 1.29, en el segundo 2.03, en el tercero 1.56 y en el cuarto 1.63. Notamos que, en todos los años, el componente E_tI_t del cuarto trimestre tiene una influencia por encima del promedio, es decir, es mayor que 1.00. Además, identificamos un valor relativamente alto (2.03).

El efecto que combina E_tI_t puede ser refinado eliminando las fluctuaciones irregulares de los datos, permitiendo la determinación de índices de estacionalidad. El cuadro 24 presenta los

valores de $E_t I_t$ en una tabla de doble entrada por años y trimestres. Cada columna representa el efecto combinado de las influencias estacionales e irregulares en un trimestre determinado. Las diferencias entre los valores de cada trimestre se atribuyen a las fluctuaciones irregulares; cuando estas son fuertes se presentan valores extremos –muy altos o muy bajos–, que deben ser ignorados para evitar sesgos en la estimación de los factores estacionales. Eliminando estos valores extremos y calculando el promedio de los valores restantes obtenemos una primera aproximación del factor de estacionalidad para cada trimestre. Así, para el primer trimestre eliminamos el valor correspondiente al tercer año, por ser muy bajo (B), y calculamos el promedio de los otros tres valores (0.80). Del mismo modo se calcula el valor 0.63 para el segundo trimestre, 0.97 para el tercero y 1.49 para el cuarto.

Un último ajuste es necesario para obtener los *factores de estacionalidad*. El modelo multiplicativo requiere que el factor de estacionalidad promedio sea igual a 1.00; es decir, que la suma de los cuatro factores debe ser igual a 4.00. Esto es necesario si los efectos estacionales se han de prorratear a través del año. Dado que, en nuestro ejemplo, la suma de los promedios sin irregularidades es de 3.89, calculamos un factor de ajuste de la siguiente manera:

$$\text{Factor de ajuste} = \frac{4.00}{3.89} = 1.028$$

Al multiplicar los promedios por el factor de ajuste obtenemos los factores de estacionalidad, que suman 4.00. En el cuadro 24 se resumen todos los pasos necesarios para calcular los factores de estacionalidad para las ventas de órganos electrónicos. También se muestran los índices de estacionalidad que se obtienen multiplicando los factores de estacionalidad por 100.

CUADRO 24: CÁLCULO DE LOS NÚMEROS ÍNDICES ESTACIONALES

Año	Primer trimestre	Segundo trimestre	Tercer trimestre	Cuarto trimestre
1			1.23 A	1.29
2	0.890	0.68	0.87	2.03 A
3	0.40B	0.61	1.08	1.56
4	0.840	0.60	0.95	1.63
5	0.670	0.75 A		
Promedio sin irregularidades	0.800	0.630	0.97	1.49
Promedio por ajuste* (factor estacional)	0.822	0.648	0.997	1.532
Índice de estacionalidad	82.200	64.8	99.7	153.2

* Suma de promedios = 3.89. Factor de ajuste = $\frac{4.00}{3.89}$ = 1.028

Los índices de estacionalidad nos brindan información numérica sobre la influencia de las fluctuaciones estacionales en la serie histórica de ventas de órganos electrónicos. El cuarto trimestre de cada año es el de mejores ventas, con un nivel de 53% por encima del promedio trimestral. El trimestre de menores ventas es el segundo, con un índice de estacionalidad de 64.8%, lo que indica que las ventas promedio para este trimestre están 35.2% por debajo de las ventas trimestrales promedio.

B. Desestacionalización para encontrar el patrón de tendencia

Para identificar la tendencia en una serie de tiempo que contiene efectos estacionales debemos primero remover el efecto estacio-

nal. Este proceso se conoce como *desestacionalización de la serie de tiempo*, y consiste en dividir cada observación histórica, Y_t, por el factor de estacionalidad correspondiente. De esta manera identificamos el efecto combinado de tendencia, cíclico e irregular:

$$\frac{Y_t}{E_t} = \frac{T_t * C_t * E_t * I_t}{E_t} \quad (19)$$

Los valores desestacionalizados de nuestro ejemplo se presentan en el cuadro 25 y se ilustran en el gráfico 47. En este gráfico vemos que, a pesar de movimientos hacia arriba y hacia abajo en los últimos veinte trimestres, la serie de tiempo parece tener una tendencia creciente. Podríamos ajustar una línea recta o curva a este patrón de tendencia. Para identificar esta tendencia usaremos el mismo procedimiento utilizado para determinar la tendencia de una serie con datos anuales.

Si suponemos una tendencia lineal, entonces el volumen de ventas estimadas puede expresarse como una función del tiempo:

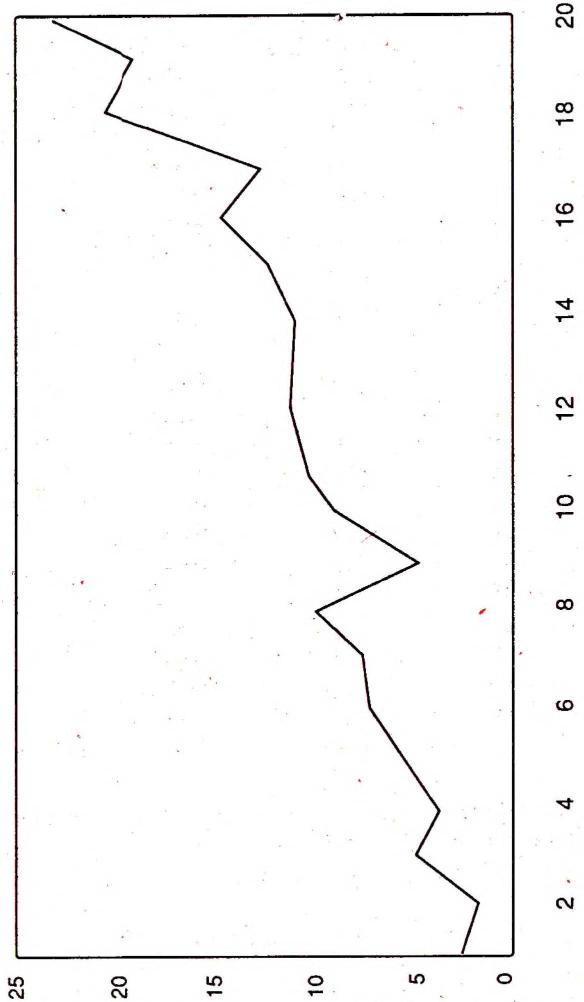
$$T_t = a + bt \quad (20)$$

donde T_t es el valor de tendencia para las ventas en el período t.
a es el intercepto de la línea de tendencia.
b es la pendiente de la línea de tendencia.
t es el período trimestral (t = 1,2,3, ...,20).

Usando las fórmulas del modelo de regresión lineal presentadas en el capítulo anterior, ecuaciones (12) y (13), calculamos los valores de a y b:

$$a = -0.3201 \qquad b= 0.9379$$

Gráfico 47: Ventas desestacionalizadas de órganos electrónicos en los últimos veinte trimestres



CUADRO 25: VENTAS DESESTACIONALIZADAS

Año/Trimestre		Ventas ($Y_t = T_tC_tE_tI_t$)	Factor estacional (E_t)	Ventas desestacionalizadas ($Y_t/E_t = T_tC_tI_t$)
1	I	2	0.822	2.43
	II	1	0.648	1.54
	III	4	0.997	4.01
	IV	5	1.532	3.26
2	I	4	0.822	4.87
	II	4	0.648	6.17
	III	6	0.997	6.02
	IV	14	1.532	9.14
3	I	3	0.822	3.65
	II	5	0.648	7.71
	III	10	0.997	10.03
	IV	16	1.532	10.44
4	I	9	0.822	10.95
	II	7	0.648	10.80
	III	12	0.997	12.04
	IV	22	1.532	14.36
5	I	10	0.822	12.17
	II	13	0.648	20.06
	III	18	0.997	18.05
	IV	35	1.532	22.85

Reemplazando estos valores en la ecuación (20), obtenemos el componente de tendencia lineal de la serie de tiempo de ventas de órganos electrónicos:

$$T_t = -0.3201 + 0.9379t \tag{21}$$

La pendiente 0.94 nos indica que en los últimos veinte trimestres las ventas de la empresa han experimentado un crecimiento

promedio desestacionalizado de aproximadamente 0.94 órganos electrónicos por trimestre. Si *suponemos* que la tendencia de las ventas en los últimos veinte trimestres seguirá vigente en el futuro, entonces podemos utilizar la ecuación (21) para predecir el componente de tendencia de las ventas de órganos en trimestres futuros.

C. Las fluctuaciones cíclicas

El último y menos significativo de los componentes sistemáticos que afectan una serie de tiempo es el que corresponde a las variaciones cíclicas. Algunos economistas consideran que las actividades económicas sufren cierto tipo de movimiento oscilatorio cada cierto número de años. Estos ciclos comprenden cuatro fases: depresión, recuperación, prosperidad y recesión. La bibliografía sobre ciclos es amplia; sin embargo, no hay consenso acerca de su naturaleza, causas o, incluso, de la existencia misma del ciclo.

El componente cíclico en una serie de tiempo podría identificarse una vez desestacionalizada la serie, $T_tC_tI_t$, y estimada la tendencia, T_t:

$$\frac{\text{Serie desestacionalizada}}{\text{Tendencia}} = \frac{T_tC_tI_t}{T_t} = C_tI_t$$

Sin embargo, estos valores todavía incluirían algunas fluctuaciones irregulares, sin existir una técnica adecuada para aislarlos. Además, para poder identificar los movimientos oscilatorios y estudiar el componente cíclico de manera adecuada sería necesario contar por lo menos con quince años de observaciones históricas.

D. Recomposición y predicción

Una vez identificados los patrones de tendencia (T_t) y estacional (E_t) usando el modelo multiplicativo, intentaremos formular

predicciones acerca de las ventas de órganos electrónicos en el futuro. Esto será posible siempre y cuando supongamos que los factores de comportamiento de estas ventas en los últimos veinte trimestres continuarán en el futuro. Primero proyectamos la tendencia para los trimestres futuros y luego ajustamos esta predicción usando los factores de estacionalidad respectivos.

La tendencia de las ventas se proyecta con base en la ecuación (21). Así, la tendencia para el primer trimestre del sexto año, es decir el trimestre 21, se obtiene sustituyendo t = 21 en dicha ecuación:

$$T_{21} = -0.3201 + 0.94*21 = 19.42$$

Considerando únicamente el componente de tendencia, pronosticamos ventas de 19.42 órganos electrónicos para el próximo trimestre. De igual manera, proyectaremos ventas de 20.36, 21.30 y 22.24 órganos en los trimestres 22, 23 y 24 respectivamente.

Ahora debemos ajustar estas predicciones para tomar en cuenta el efecto estacional. Así, la predicción para el primer trimestre del sexto año se obtiene multiplicando la predicción tendencial por el factor de estacionalidad correspondiente al primer trimestre. Y así con el resto de predicciones tendenciales. Los valores ajustados se presentan en el cuadro 26.

CUADRO 26: PREDICCIONES TRIMESTRALES DE LAS VENTAS DE ÓRGANOS ELECTRÓNICOS

Año/Trimestre	Predicción tendencial (T_t)	Factor estacional (E_t)	Predicción trimestral (T_tE_t)
6.I	19.42	0.822	15.96
6.II	20.36	0.648	13.19
6.III	21.30	0.997	21.24
6.IV	22.24	1.532	34.07

Estas predicciones se podrían ajustar aún más considerando los efectos cíclicos, análisis que no estamos en capacidad de realizar por no contar con suficiente información para identificar las fluctuaciones cíclicas.

4. PREDICCIONES CON MÉTODOS CAUSALES

Mientras que las predicciones que se efectúan utilizando los métodos de series de tiempo se basan en el análisis del comportamiento de una sola variable a través del tiempo, las predicciones con métodos causales toman en consideración el análisis de más de una variable. Estos últimos métodos estudian las relaciones entre las variables identificando los factores causales para poder proyectar el comportamiento de las variables de interés. Los modelos causales de predicción más comunes son el análisis de regresión, el análisis de insumo-producto y los modelos econométricos. Todos ellos pueden ser empleados a diferentes niveles de agregación: desde el nivel empresarial hasta la planificación de la economía nacional.

A. Análisis de regresión

Un modelo de regresión considera una o más variables independientes que explican el comportamiento de otra, llamada variable dependiente. El análisis de regresión y la estimación de modelos específicos utilizando la técnica de mínimos cuadrados se presenta de manera extensa en el capítulo VI.

B. Modelo de insumo-producto

Es un modelo descriptivo de la economía, que muestra el volumen de producción de cada sector y los tipos y cantidades de insumos que se han utilizado para obtener dicha producción. Su utilidad consiste en la posibilidad de predecir, con base en

este modelo, las consecuencias de determinadas políticas económicas en términos de producción sectorial. Por ejemplo, una política de promoción de exportaciones del sector j que incremente la producción de este sector, desencadenará en forma directa una mayor producción en todos aquellos sectores que producen insumos para el sector j. Asimismo, se generarán sucesivos incrementos de producción en otros sectores que producen insumos para los sectores afectados directamente.

El modelo de insumo-producto también puede utilizarse para evaluar las políticas de precios, pues sirve para medir los efectos de la variación de ciertos precios en el nivel general de inflación. Por ejemplo, un incremento en el precio de los combustibles generará variaciones en el precio de los otros bienes, dadas las relaciones interindustriales.

C. Modelos econométricos

Un modelo econométrico describe el funcionamiento de un sistema económico a través de un conjunto de ecuaciones simultáneas. La formulación y estimación de dichas ecuaciones se basa en la teoría económica, estadística, matemática y la contabilidad, para describir la interdependencia entre las diferentes partes del sistema económico. Este sistema puede referirse a la economía global de un país o a una empresa individual.

Los coeficientes de las ecuaciones de los modelos econométricos se estiman simultáneamente utilizando técnicas más sofisticadas que la de mínimos cuadrados ordinarios, tales como los métodos de variables instrumentales, mínimos cuadrados bietápicos, trietápicos, etcétera. Estas estimaciones requieren de cuantiosa información histórica y facilidades de procesamiento electrónico.

5. PREDICCIONES CUALITATIVAS

En las secciones anteriores hemos discutido varios tipos de métodos cuantitativos de predicción que requieren de datos históricos. Además, estos métodos suponen que las condiciones del entorno que afectan el comportamiento de las series de tiempo se mantendrán vigentes durante el período de predicción.

En situaciones en las que no se dispone de datos históricos y/o se producen cambios que alteran las condiciones pasadas, será necesario utilizar técnicas alternativas. Las técnicas de predicción cualitativas tratan de coordinar en una forma no-sesgada y sistemática todas las opiniones o juicios relevantes para producir una predicción.

A. Predicciones de una persona

En muchas organizaciones, y particularmente en pequeñas empresas, la predicción de una variable es tarea de una persona. Por ejemplo, un pequeño agricultor establece su plan de producción para la próxima campaña sobre la base de su predicción de los precios de los diferentes productos que puede sembrar. Dicha predicción es usualmente una adivinanza basada en los precios actuales. En algunos casos esta predicción personal se enriquece con consultas a otros agricultores o con el apoyo de asesores.

B. Paneles

Este método consiste en reunir a un grupo de expertos para intercambiar opiniones, complementar conocimientos y efectuar, en conjunto, una predicción adecuada de la variable de interés. La base de este método es que cada experto reconoce la capacidad de los otros en su área y que el trabajo en equipo logrará una mejor predicción.

La ventaja del método de paneles es que las predicciones pueden obtenerse inmediatamente. Sin embargo existen limitaciones, dado que el método se basa únicamente en las experiencias presentes y cada experto podría opinar tomando como base diferentes supuestos no explicitados. Además, existen ciertos factores sociales relacionados con la interacción en grupos que pueden llevar a predicciones sesgadas. Frecuentemente se observa que los participantes se adhieren a la opinión de la mayoría o a la opinión de la persona de mayor jerarquía en el grupo, aunque sus propias opiniones sean de más valor que las emitidas por sus superiores. El método Delfi intenta superar algunas de las limitaciones originadas por la interacción social.

C. El método Delfi

Este método utiliza una secuencia de cuestionarios para recoger e intercambiar las opiniones de un conjunto de expertos sobre el pronóstico deseado. Al igual que en los paneles, el objetivo es lograr una predicción por consenso grupal, pero la diferencia es que los expertos que participan en la tarea de predicción están físicamente separados unos de otros y sus opiniones individuales permanecen anónimas.

Un primer cuestionario es enviado a cada miembro del panel, y sus respuestas se utilizan para generar un segundo cuestionario. Las primeras respuestas permiten calcular el valor promedio del pronóstico y el rango intercuartil (RIQ). Un segundo cuestionario, también enviado a los panelistas, les comunica los resultados del primero (media y RIQ) y les pide que revisen sus pronósticos. Si desean mantener su pronóstico fuera del RIQ, se les solicita que establezcan sus razones. Las nuevas respuestas se usan a su vez para generar un tercer cuestionario; se vuelve a calcular un nuevo promedio y RIQ. Estos valores, así como las razones para los puntos fuera del RIQ, se incluyen en el tercer cuestionario, y una vez más se les pide a los panelistas

que elaboren un pronóstico revisado a la luz de la nueva información.

Este proceso continúa hasta que los expertos que coordinan la predicción tienen información suficiente y consideran que se ha logrado algún grado de consenso. Vemos, pues, que el objetivo del método Delfi no es dar una respuesta única como resultado final, sino más bien producir un rango pequeño de opiniones con el que está de acuerdo la mayor parte de los expertos del panel.

Ejercicios

1. ¿Cuál de los componentes de las series de tiempo esperaría usted encontrar en cada una de las siguientes series?
a. Las ventas mensuales de helados en la ciudad de Lima en el período de enero de 1980 a enero de 1991.
b. La producción trimestral de cebollas en el departamento de Arequipa de 1985 a 1991.
c. Las ventas anuales de una compañía de seguros de 1985 a 1991.
d. El número de atentados terroristas mensuales en Lima Metropolitana de 1982 a 1991.

2. Sugiera un método de predicción a ser utilizado en cada uno de los siguientes problemas. Argumente su selección y defina cuáles serían los datos relevantes para sus pronósticos.
a. El crecimiento del sector construcción a partir del próximo año.
b. La demanda de energía eléctrica, para cada uno de los próximos cinco años, para uso comercial y doméstico en las ciudades de Lima y Tacna, y en todo el país.
c. Las ventas de pólizas de seguros de vida para el próximo año en todo el país.
d. La tasa de desempleo en el Perú para cada uno de los siguientes doce meses.
e. Las ventas de bebidas no alcohólicas en la ciudad de Trujillo para cada uno de los siguientes cuatro trimestres.

3. El número de alumnos matriculados en una academia de preparación preuniversitaria, en la ciudad de Lima, durante los últimos seis años, se presenta en el siguiente cuadro:

Año	Alumnos matriculados
1	2,050
2	2,020
3	1,950
4	1,900
5	1,910
6	1,880

a. Grafique la serie de los alumnos matriculados con respecto al tiempo.
b. Estime el número de alumnos a matricularse en el año 7, usando:
(i) una predicción histórica
(ii) la media aritmética
(iii) una media móvil adecuada
(iv) la técnica de suavización exponencial
c. En su opinión, ¿cuál de estas predicciones es más adecuada?

4. Refiérase al problema anterior. Establezca una expresión para identificar si existe una tendencia lineal en el número de estudiantes matriculados.
a. ¿Es el ajuste de tendencia lineal adecuado para la serie bajo análisis?
b. Grafique los valores observados y los valores ajustados.
c. Comente sobre el número de alumnos matriculados en esta academia en los últimos seis años.
d. Estime el número de alumnos matriculados para el año 7.
e. Compare las predicciones efectuadas en el problema anterior con la predicción que usa la tendencia lineal. ¿Qué conclusión puede extraer?

5. "Radios S.A." es una empresa que se dedica a la comercialización de aparatos receptores-transmisores para radioaficionados. El número de equipos vendidos en sus siete años de negocio es el siguiente:

Año	1	2	3	4	5	6	7
Ventas	35	50	75	90	105	110	130

a. Elabore un gráfico para esta serie de tiempo.
b. Ajuste una tendencia lineal a esta serie. ¿Es este un buen ajuste? Explique.

c. Use la tendencia lineal desarrollada en (b) y prepare una proyección para las ventas anuales en el año 8.

6. Suponga que la tendencia de las ventas de pastas de la tienda "Aurelia" es la siguiente:

$$Y = 300 + 100\,X$$

El origen de la serie de tiempo es enero de 1989; X es un intervalo de un mes, e Y representa el número de kilos vendidos.

a. ¿Cuál es la predicción de ventas para enero de 1992, junio de 1992 y enero de 1993?

b. Se encontró que las ventas de pastas son estacionales. Los índices de estacionalidad son 175 para junio y 95 para enero. Ajuste sus predicciones de la parte (a).

7. Los datos para las ventas trimestrales de textos universitarios en los últimos tres años son los siguientes:

	Año 1	Año 2	Año 3
Trimestre 1	940	900	1,100
Trimestre 2	2,625	2,900	2,930
Trimestre 3	2,500	2,360	2,615
Trimestre 4	1,690	1,800	1,850

a. Grafique la serie de ventas de textos con respecto al tiempo.

b. ¿Cuáles de los componentes parecen estar presentes en el patrón de ventas?

c. Calcule los factores o índices estacionales para los cuatro trimestres.

d. ¿Cuándo es que la venta de textos experimenta el efecto estacional más grande? ¿Es esto razonable? Explique.

8. Usando los resultados del problema anterior,

a. Desestacionalice la serie de ventas.

b. Grafique los valores observados y los valores desestacionalizados.

c. Ajuste una tendencia lineal a la serie desestacionalizada.

d. Usando los resultados anteriores, estime las ventas trimestrales para el año 4.

9. El Gerente de un hotel de turismo utilizó los datos referentes a seis años para calcular un índice de variación estacional para los gastos de turistas en su hotel, obteniendo el siguiente resultado:

Mes	Índice
Enero	125
Febrero	155
Marzo	100
Abril	40
Mayo	50
Junio	115
Julio	120
Agosto	135
Setiembre	60
Octubre	70
Noviembre	110
Diciembre	120

Si se proyecta que el gasto total de los turistas en el hotel será de $ 36,000 en 1992, estime el gasto esperado para cada mes.

10. A continuación se presentan (en miles de dólares) las ventas trimestrales de una fábrica textil en los últimos tres años:

	Año 1	Año 2	Año 3
Trimestre 1	250	236	261
Trimestre 2	262	290	293
Trimestre 3	169	180	185
Trimestre 4	94	90	110

a. Grafique la serie de las ventas trimestrales con respecto al tiempo.
b. Calcule la media móvil de cuatro trimestres para esta serie de tiempo.
c. Calcule los factores estacionales para los cuatro trimestres.
d. ¿Cuándo ocurre el efecto estacional más grande?

11. Con los datos del problema anterior:

a. Desestacionalice la serie de las ventas, usando los factores de estacionalidad allí calculados.
b. Grafique los valores desestacionalizados. ¿Existe alguna tendencia?

c. Ajuste una tendencia lineal a la serie desestacionalizada.
d. Usando los patrones de tendencia y de estacionalidad identificados, formule predicciones de las ventas trimestrales para el año 4.

12. Un Gerente a cargo del control de inventarios requiere predicciones de las ventas de varios productos para los próximos seis meses. Este Gerente cuenta con datos históricos mensuales de los últimos cuatro años. Decide usar como predicción para cada uno de los próximos seis meses el promedio mensual en los últimos cuatro años. ¿Piensa usted que esta es una buena estrategia? Explique su razonamiento.

VIII. Teoría de decisiones

1. Las bases fundamentales de toda decisión. 2. Árboles de decisiones. 3. Elección de la alternativa óptima. 4. Análisis de sensibilidad. 5. Valor esperado de la información perfecta. 6. Toma de decisiones usando información muestral. 7. Estrategia óptima de decisión con información muestral. 8. Valor esperado de la información muestral.

La estadística inferencial y el análisis de regresión, vistos en los capítulos anteriores, se basan por completo en el uso de información muestral. Tanto en el problema de *estimar* alguna característica de la población como en el de *probar una hipótesis* en relación a algunas características de la población, el método utilizado es siempre el mismo. Se selecciona una muestra aleatoria de la población y la información muestral sirve de base para todas las inferencias que se hagan en relación a la población. En este proceso inferencial se presta poca atención a la toma de decisiones; no le preocupan los problemas empresariales en los que la experimentación formal juega tan sólo un papel secundario. Esto se debe a que el curso de acción que se tome después de hacer una inferencia estadística depende comúnmente de consideraciones externas, ajenas al estadístico.

El *enfoque clásico o empírico* del análisis estadístico tiene por objetivo *inferir* sobre ciertas características poblacionales, basándose en la información contenida en una muestra de la población. Sin embargo, en economía y administración existen ciertos

tipos de problemas en los que no es posible obtener muestras para estimar ciertas características de la población. Es necesario recurrir a información subjetiva o información de una persona. En general, podemos decir que la información que puede utilizarse para hacer una inferencia en relación a una característica poblacional se clasifica en dos tipos: objetiva y subjetiva. Así, la información obtenida mediante un muestreo es información objetiva. La opinión de un experto en la materia, por otra parte, es información subjetiva.

El enfoque clásico, que utiliza información muestral objetiva, fue extendido significativamente durante la década de los 50 y permitió utilizar toda la información relevante: objetiva y subjetiva. Este enfoque se conoce como la *teoría de decisiones*, y permite al decisor integrar al proceso mismo de toma de decisiones sus percepciones sobre incertidumbre, su ingenio para crear alternativas de acción y sus apreciaciones personales de la estructura del problema. La teoría de decisiones podría definirse como el análisis lógico y cuantitativo de todos los factores que afectan los resultados de una decisión en un mundo incierto. Esta teoría prescribe cómo debe proceder un individuo que se enfrenta a un problema de elección con incertidumbre para escoger un curso de acción coherente con la información disponible y sus preferencias personales. Pero este enfoque estuvo recluido al mundo académico, para analizar problemas simples pero inciertos. Sin embargo, hubo un desarrollo teórico suficiente como para responder a la pregunta: ¿qué es una buena decisión?

En la siguiente sección se establecerá la diferencia entre la calidad de una decisión y de su resultado, para luego señalar las bases fundamentales de toda decisión. En la sección 2 se mostrará cómo el proceso de toma de decisiones puede ser estructurado a través de "árboles de decisiones". La solución del árbol de decisiones, usando el valor esperado como criterio de decisión, se presenta en la sección 3. El cálculo del valor esperado de

las diferentes alternativas depende de las probabilidades que se usan. Estas probabilidades son evaluadas sobre la base de experiencias previas del decisor o sus expertos. En la sección 4, análisis de sensibilidad, buscaremos responder a la siguiente pregunta: ¿entre qué límites pueden variar las probabilidades evaluadas sin alterar la decisión óptima? En la sección 5 se introducirá el concepto del valor esperado de la información perfecta, y se discutirá la manera de calcularlo. Finalmente, en las últimas secciones veremos cómo utilizar información adicional, que no es perfecta, en el proceso de toma de decisiones, y cómo determinar su valor para el decisor.

1. LAS BASES FUNDAMENTALES DE TODA DECISIÓN

El propósito de la teoría de decisiones es incrementar la probabilidad de obtener buenos resultados en un mundo de incertidumbres. Un *buen resultado* es aquel que es apreciado favorablemente por el decisor; es decir, aquel que le gustaría que ocurriera. Por otro lado, una *buena decisión* es aquella a la que se llega considerando integral, lógica y explícitamente la información, alternativas y preferencias del decisor.

Una *decisión* es una asignación *irreversible* de recursos, en el sentido que para cambiar la decisión será necesario invertir recursos adicionales que pudieran resultar prohibitivos. Hay decisiones *inherentemente* irreversibles, como la amputación de la pierna de un paciente. Hay otras *esencialmente* irreversibles, como la decisión de lanzar un nuevo producto al mercado. Entonces, es claro que para tomar una decisión es imprescindible contar con los recursos necesarios. Tener interés en un problema sin disponer de los recursos, y por ende sin la capacidad de tomar decisiones, es simplemente una "preocupación" y nada más.

El "decisor" es el individuo, o conjunto de individuos, que tiene la responsabilidad de comprometer o asignar recursos de

una organización. El decisor puede ser el Jefe de Almacén, el Gerente de Producción, el Gerente General o el Directorio de una empresa, dependiendo del nivel de la decisión. En un problema de decisiones es crucial identificar claramente al decisor, pues la calidad de la decisión dependerá de si esta es o no consistente con las alternativas, información y preferencias de quien en última instancia es responsable por la decisión.

En el proceso de toma de decisiones nos enfrentamos a un medio ambiente incierto, complejo, dinámico, competitivo y finito, lo que produce una reacción humana de confusión y preocupación. Pero el hombre tiene «armas» disponibles para enfrentarse a este medio ambiente. Estas pueden clasificarse en tres tipos: ingenio, percepción y filosofía.

A través de su *ingenio* el hombre concibe y formula diferentes cursos de acción; define *alternativas* potenciales. Es a través de su ingenio que puede contestar a la pregunta: ¿qué puedo hacer? Así, se podrá enumerar una lista de alternativas específicas que estarán disponibles para solucionar su problema de decisión.

En segundo lugar, a través de su *percepción* el hombre puede aprender de lo que ve, formándose juicios acerca del medio ambiente. Es decir, el hombre puede acumular *información* de su medio ambiente. Este cúmulo de información permite realizar dos tareas básicas: caracterizar las incertidumbres por medio de la asignación de probabilidades, y representar las relaciones entre los factores que influyen sobre los resultados de una decisión. Entre estos factores existen algunos que están fuera del control del decisor. Estos eventos futuros inciertos son conocidos como *variables de estado* del problema. Para cada una de ellas se tendrá una lista de todos los posibles valores que pueda tomar, llamándoseles *estados de la naturaleza o del entorno*. Sólo uno de ellos ocurrirá, indicando que los valores de la variable de estado son eventos mutuamente excluyentes y colectivamente exhaustivos. Entonces podemos contestar a la pregunta básica: ¿qué sabemos?

Finalmente, y quizá lo más importante, el hombre tiene una *filosofía;* principios que guían su vida y que definen sus *preferencias* respecto a los varios resultados que puede obtener de su decisión. Podrá contestar a la pregunta: ¿qué quiero? Las preferencias se dividen en tres categorías. Primero, las *preferencias con respecto a resultados conflictivos.* ¿Cuánto más vale este resultado con respecto a aquel?; ¿cuál es el valor relativo de los dos? Debemos asignar valores a estos resultados para poder hacer comparaciones. En problemas económico-financieros, muchas veces todos los resultados podrán expresarse en términos monetarios, simplificando las comparaciones relativas. Segundo, las *preferencias con respecto al tiempo.* Estas preferencias se refieren al valor que se asigna a resultados que están distribuidos a través del tiempo. ¿Estamos dispuestos a aceptar un menor beneficio si podemos obtenerlo más pronto? Finalmente, tenemos las *preferencias con respecto al riesgo.* Es un término que usamos para describir el hecho de que la mayoría de las personas no son decisores neutrales al riesgo que usan el valor esperado para escoger entre diferentes alternativas. Por ejemplo, ¿cuántos de nosotros estaríamos dispuestos a apostar en el lanzamiento de una moneda a duplicar nuestros ingresos del próximo mes o no recibir nada?

En resumen, los tres elementos básicos que interactúan en todo proceso de toma de decisiones son: las alternativas, la información y las preferencias. Cuando hayamos definido las *preferencias,* cuando hayamos establecido los *modelos necesarios* para evaluar la decisión que estamos considerando y cuando hayamos *asignado probabilidades* a las variables inciertas, entonces la elección de la mejor alternativa seguirá de una manera lógica y directa. Por tanto, es fácil definir una *buena decisión* como aquella que resulta como una consecuencia lógica del análisis de las alternativas, de la sistematización de la información relevante y de las preferencias explícitas del decisor. La disciplina del *análisis de decisiones* ha desarrollado metodologías adecuadas

para evaluar preferencias (al riesgo, tiempo y resultados diferentes), así como para asignar probabilidades y estructurar o modelar el problema convenientemente.

El presente capítulo tiene como objetivo hacer una exposición básica de la *teoría de decisiones*, que es uno de los soportes conceptuales de la disciplina del análisis de decisiones. Con ese fin, es necesario establecer algunos supuestos para simplificar el análisis. En primer lugar, supondremos que todos los resultados serán obtenidos de inmediato y por lo tanto no será necesario evaluar las preferencias con respecto al tiempo. En segundo lugar, todos los resultados posibles se expresan en unidades monetarias, lo que hará más fácil la comparación de posibles resultados conflictivos. Finalmente, respecto a las preferencias al riesgo, supondremos que el decisor es neutral al riesgo; es decir, que está dispuesto a aceptar el valor esperado como criterio de decisión.

2. ÁRBOLES DE DECISIONES

El árbol de decisiones es una técnica que se utiliza para estructurar el proceso de toma de decisiones bajo incertidumbre, y su análisis se basa en la teoría de probabilidades. Para ilustrar el uso del árbol de decisiones, consideremos el caso de la Compañía de Manufacturas Eléctricas (CME), que produce aparatos de aire acondicionado. CME debe decidir si comprar un componente importante para su producto final de un abastecedor, o fabricarlo en su propia planta.

En el análisis de toda decisión, el primer paso es responder a la pregunta ¿qué podemos hacer?, para identificar las *alternativas* disponibles que puede considerar el decisor. En el caso sencillo de CME, las alternativas son obvias:

C: Compra el componente.
F: Fabrica el componente.

La determinación de la mejor decisión dependerá de la aceptación de su producto final en el mercado y, consecuentemente, de la demanda derivada del componente en cuestión. Dado que la demanda que CME enfrenta por su producto final está fuera del control del decisor, esta constituye una *variable de estado*. De acuerdo con la administración de CME, los posibles valores de la demanda por su producto final pueden ser altos, medios o bajos. Luego, los estados de la naturaleza que CME podría enfrentar son:

DA = Demanda alta del producto final de CME.
DM = Demanda media del producto final de CME.
DB = Demanda baja del producto final de CME.

Dadas las alternativas de decisión y los tres posibles estados de la naturaleza, ¿cuál es la decisión óptima para CME? Para contestar a esta pregunta necesitamos mayor *información* respecto a las *probabilidades* de ocurrencia de cada posible estado de la naturaleza, y respecto a la relación que existe entre las variables de decisión y de estado que influyen en los resultados de la decisión. En el caso de CME la estructuración del problema es muy simple, dado que sólo tenemos una variable de decisión (compra/fabricación del componente) y una variable de estado (nivel de demanda). Debemos buscar información respecto a las ganancias netas asociadas con cada combinación de una alternativa de decisión y un estado de la naturaleza. Por ejemplo, ¿cuál es la ganancia que CME obtendrá si decide comprar el componente y la demanda fuera alta?; ¿cuál es la ganancia si la empresa decide producir el componente y la demanda es media?; y así sucesivamente.

El *resultado final* de la decisión se expresa en términos de ganancias netas; es decir, en una medida de rentabilidad. Utilizando la mejor información disponible, la administración de CME ha estimado las ganancias netas para este problema, las

cuales se muestran en el cuadro 27. Este cuadro, que se conoce como una *tabla de resultados*, muestra las ganancias netas para cada combinación posible de las variables de decisión y de estado. En general, los contenidos de una tabla de resultados pueden expresarse en términos de ganancias netas, costos o cualquier otra medida adecuada que represente el resultado de la decisión en la situación particular que se está analizando.

CUADRO 27: TABLA DE RESULTADOS DEL PROBLEMA DE CME. GANANCIAS NETAS CONDICIONALES
(Miles de dólares)

Alternativas de decisión	Estados de la natulareza (Niveles de demanda)		
	Demanda alta (DA)	Demanda media (DM)	Demanda baja (DB)
Fabricación del componente (F)	130	40	-20
Compra del componente (C)	70	45	10

Si la demanda por el producto final es alta, será ventajoso para CME fabricar el componente, obteniendo una ganancia neta de US$ 130,000. En cambio, si CME decide fabricar el componente y la demanda es baja, los costos unitarios de producción serán altos debido a la subutilización del equipo de producción; en este caso los ingresos no compensarán los costos, generando una pérdida neta de US$ 20,000. Por otro lado, si CME decide comprar el componente y la demanda es alta, sus ganancias serán únicamente US$ 70,000, y si la demanda es baja las ganancias serán positivas e iguales a US$ 10,000.

Tal como se indica en el encabezamiento del cuadro 27, estas ganancias se denominan ganancias condicionales, ya que cada

una depende de –o está condicionada por– la decisión de fabricar/producir el componente y del nivel de demanda que enfrente CME por su producto final. Puesto que el nivel de demanda es una variable de estado, está *fuera del control del decisor*. Los problemas de decisión, en general, se caracterizan por la presencia de incertidumbre en relación a la ocurrencia de los diversos sucesos o estados de la naturaleza. Si bien estos sucesos están fuera del control del decisor, este puede asignar probabilidades subjetivas a cada uno de los posibles valores de las variables de estado *basándose en la información disponible*. Esta información es producto de la experiencia, juicio y expectativas del decisor y/o sus expertos. Estas probabilidades subjetivas se denominan *probabilidades a priori* y están sujetas a revisión a la luz de cualquier información adicional. En el caso de CME, el decisor debe asignar probabilidades *a priori* a los diversos niveles de demanda de su producto final.

De acuerdo con la experiencia de la administración de CME, se asignó una probabilidad de ocurrencia de 0.30 al nivel alto de demanda, 0.30 a la demanda media y 0.40 a la probabilidad de una demanda baja por el producto final.

Una vez que los elementos básicos del problema de decisión han quedado establecidos, el decisor está en condiciones de elegir la mejor alternativa. Puesto que suponemos que nuestro decisor es neutral al riesgo, este evaluará y comparará el valor esperado de cada acción alternativa expresada en ganancias netas, y elegirá aquella que le represente el mayor beneficio.

Los problemas de decisión que involucran una sola variable de decisión y una variable de estado pueden ser analizados utilizando las tablas de resultados. Pero cuando los problemas son más complejos y el número de estas variables aumenta, la construcción de tales tablas se torna imposible. Para estos casos, así como para los casos más simples, podemos usar el diagrama del *árbol de decisiones*.

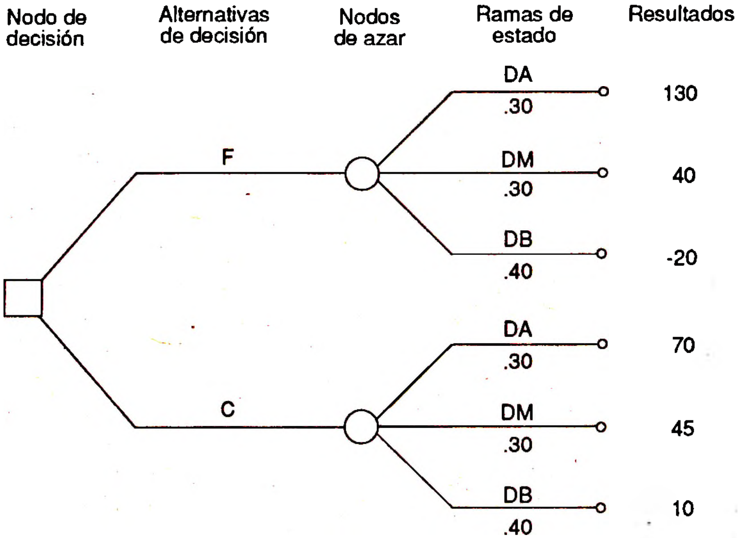
En general, un árbol de decisiones se construye horizontalmente comenzando del extremo izquierdo, como se hizo con los árboles de probabilidades. El árbol de decisiones muestra la progresión natural o lógica que ocurre en el proceso de toma de decisiones. En primer lugar, para cada variable de decisión colocamos un *nodo cuadrado*, del cual saldrán tantas líneas o ramas como alternativas disponibles existan en dicha decisión. Segundo, cada rama es denotada por la alternativa disponible respectiva. Tercero, los terminales de las ramas de decisión son usados como nodos de comienzo de *variables de estado*. Se usan nodos redondos para definir las diferentes variables de estado del problema bajo análisis. De cada uno de estos nodos salen tantas ramas como posibles valores pueda tomar la primera variable de estado en el problema. Cuarto, los valores de la variable de estado se consignan encima de la rama respectiva, y su probabilidad de ocurrencia debajo de la misma. El proceso de construcción continúa hasta que todas las posibles secuencias de variables de decisión y variables de estado han sido representadas.

Las probabilidades que se registran en las ramas de las variables de estado son probabilidades condicionales que dependen de los valores de las variables de estado y de las decisiones que las preceden. Los nodos finales del árbol representan el conjunto de todos los posibles resultados asociados con cada una de las alternativas de decisión. La probabilidad de ocurrencia de cada punto o nodo final es calculada multiplicando todas las probabilidades en la secuencia de ramas que llevan del nodo inicial del árbol al nodo final; esto es una aplicación del concepto de expansión en cadena (ver la sección 6 del capítulo II).

Una vez construido el árbol de decisiones se debe evaluar cada resultado final asociado con la secuencia respectiva de valores de las variables de estado y las variables de decisiones, para determinar las ganancias netas. En nuestro ejemplo de la CME tenemos una variable de decisión y una de estado. El

gráfico 48 muestra el árbol de este problema con la lógica progresiva que ocurre en el proceso de toma de decisiones: primero, CME debe tomar una decisión (F o C). Luego que la decisión sea ejecutada, se verificará un estado de la naturaleza expresado por el nivel de demanda. El valor en cada nodo final del árbol representa la ganancia neta asociada con cada secuencia particular de decisiones y variables de estado. Así, por ejemplo, el primer resultado de 130,000 dólares resultará cuando CME decida fabricar el componente (F) y el tamaño de mercado resulte alto (DA). El segundo resultado muestra una ganancia neta de 40,000, que se logra cuando CME toma la decisión de fabricar el componente (F) y el nivel de demanda enfrentado sea medio (DM); y así sucesivamente.

Gráfico 48: Árbol de decisiones del problema de CME



En resumen, un árbol de decisiones está compuesto por nodos de decisión, representados por cuadrados pequeños, y nodos de estado o de azar, representados por círculos pequeños. En cada nodo de decisión el decisor debe elegir una alternativa, representada por una rama de decisión. Seleccionar la mejor rama es equivalente a tomar la mejor decisión. En cambio, en los nodos de azar el decisor no puede elegir entre una u otra rama, pues estas ramas de estado son «controladas por la naturaleza».

3. ELECCIÓN DE LA ALTERNATIVA ÓPTIMA

En la sección anterior hemos representado el problema de decisión en forma de un diagrama de flujo de decisiones e incertidumbres, llamado árbol de decisiones. Este diagrama también puede ser interpretado como un mapa esquemático de carreteras, que describe en orden cronológico los movimientos que puede elegir un decisor y los movimientos gobernados por la naturaleza o el azar. Hay muchas rutas que llevan desde el nodo inicial hasta los nodos finales que representan el final de cada carretera y donde se recibirán determinados premios. El mapa, aparte de indicarnos el nombre de cada ruta, señala también las probabilidades calculadas para las posibles rutas que salen de cada nodo de azar. Dichas probabilidades son condicionales, pues dependen de la información de que se dispone en el nodo en cuestión. El problema de decisión de CME se puede expresar de la siguiente manera: Dado el árbol de decisiones o mapa de carreteras del gráfico 48, ¿cómo puede el decisor ejercer su control parcial para elegir la ruta óptima? En principio, el decisor debe elegir si quiere ir por la ruta F, que implica fabricar el componente, o por C, que significa comprarlo. Esta elección la hará usando el criterio del valor esperado (VE), dado que es un decisor neutral al riesgo.

Si el decisor elige la ruta F, el azar puede conducirlo por la ruta DA, DM o DB. ¿Cuánto vale esta opción incierta de la ruta

F? Si es lo suficientemente afortunado como para ser conducido por la ruta DA y la demanda por su producto final es alta, puede recibir 130,000 dólares; pero quizá el azar no es tan condescendiente y le conduzca por la ruta DB, sufriendo una pérdida de 20,000 dólares. De acuerdo con el criterio del VE, si el decisor elige F, sus ganancias esperadas serán:

$$\begin{aligned} VE\ (F) &= 0.30(130) + 0.30(40) + 0.40(-20) \\ &= 43 \text{ miles de dólares} \end{aligned} \tag{1}$$

Escribimos este valor esperado de F encima del nodo de azar al final de la rama F (ver gráfico 49).

De igual manera se puede calcular el VE de C.

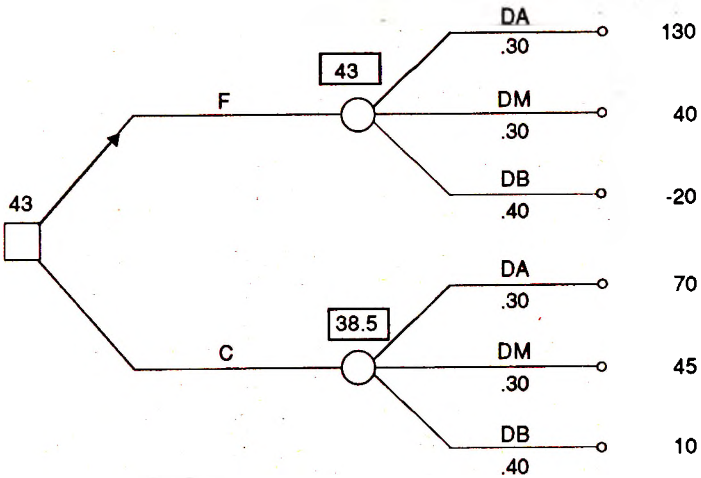
$$\begin{aligned} VE\ (C) &= 0.30(70) + 0.30(45) + 0.40(10) \\ &= 38.5 \text{ miles dólares} \end{aligned} \tag{2}$$

Una vez calculados los valores esperados en los nodos de azar, el decisor debe elegir qué ruta tomar en el nodo de decisión. ¿Qué es lo que ve por la ruta F? Una alternativa con riesgo con un valor esperado de ganancias de 43,000 dólares. ¿Qué es lo que ve por la ruta C? Una alternativa con riesgo con ganancias esperadas de 38,500 dólares. Dado que el decisor desea maximizar las ganancias netas esperadas, su elección es clara: tomará la ruta F. Esta elección se indica en el gráfico 49 con la flecha que sale del nodo de decisión. Luego, el análisis y solución del árbol de decisiones nos conduce a recomendar que CME fabrique el componente (F), pues le brindará las mayores ganancias netas esperadas: 43,000 dólares. Este valor se anota encima del nodo de decisión inicial.

Hemos visto cómo los árboles de decisiones pueden usarse para analizar un problema de decisión en un mundo simplificado pero incierto. En problemas reales de decisión, sustancialmente más complejos que el problema de CME, con muchas

variables de decisión y de estado, el enfoque del árbol de decisión es aun de mayor utilidad para lograr decisiones lógicas y consistentes con la información disponible y las preferencias establecidas.

Gráfico 49: Solución del árbol de decisiones



VE (F) = .30 (130) + .30 (40) + .40 (-20) = 43
VE (C) = .30 (70) + .30 (45) + .40 (10) = 38.5

En resumen, para analizar un problema de decisiones y establecer la alternativa óptima, el analista debe, en primer lugar, trazar el árbol de decisiones que permita estructurar cualitativamente el problema como una secuencia cronológica de elecciones que son controladas por el decisor (variables de decisión, representadas por nodos cuadrados), y de elecciones gobernadas por el azar o la naturaleza (variables de estado, representadas por nodos circulares). En segundo lugar, nos trasla-

damos conceptualmente a los extremos o nodos finales del árbol donde los resultados finales (premios, castigos, ganancias, costos u otras medidas de resultado) han sido calculados directamente de los datos del problema usando un modelo determinístico que relaciona las diferentes variables. Para resolver el árbol trabajamos hacia atrás, de derecha a izquierda, utilizando dos mecanismos: a) calculamos el valor esperado en cada nodo de azar; y, b) elegimos la mejor alternativa en cada nodo de decisión (la de máximo valor futuro si buscamos maximizar ganancias, o la de menor valor si deseamos minimizar costos).

4. ANÁLISIS DE SENSIBILIDAD

En el proceso de toma de decisiones bajo incertidumbre, las probabilidades de ocurrencia de los diferentes valores de las variables de estado de la naturaleza afectan el cálculo del valor esperado y, por tanto, la elección de la alternativa óptima. Los decisores o sus expertos evalúan estas probabilidades sobre la base de la información acumulada en experiencias previas. Sin embargo, es frecuente que en un problema de decisión se enfrenten situaciones nuevas y no se tengan experiencias previas para estimar las probabilidades relacionadas con tales circunstancias. Esta situación es típica cuando se lanza un nuevo producto al mercado y los estimados de las probabilidades correspondientes podrían ser menos confiables. En estos casos el análisis de sensibilidad respecto a las probabilidades ofrece al decisor una herramienta para determinar cuán dependiente es la decisión de los valores de las probabilidades utilizadas en el modelo.

Para el caso de CME, en el análisis de sensibilidad nos preguntamos: ¿entre qué límites puede variar la probabilidad de que la demanda sea alta (DA) sin que la decisión óptima de fabricar el componente (F) cambie? Así, si la probabilidad de enfrentar una DA fuera 0.25 en lugar de 0.30, ¿resultará todavía conveniente fabricar el componente? Al hacer este análisis de

cuán sensible es la decisión de fabricar/comprar a las variaciones de la probabilidad de DA, podremos establecer cuán importante es obtener un estimado confiable de esta probabilidad.

Usaremos el análisis de sensibilidad para establecer el rango de la probabilidad de DA en el cual la alternativa de fabricar sigue siendo la más rentable. Para determinar este rango supondremos arbitrariamente que la probabilidad de DM se mantendrá constante en 0.30 y por tanto la variación en la probabilidad de DA se verá compensada por variaciones en la probabilidad de DB. Con este supuesto podemos establecer el punto donde el valor esperado de fabricar es igual al valor esperado de comprar [VE(F) = VE(C)], es decir, aquel punto donde resulta indiferente fabricar o comprar el componente. Dado que p(DB) + p(DA) + p(DM) = 1.0 y que p(DM) = 0.30, y definiendo p(DA) = X y p(DB) = 0.70 - X, tendremos:

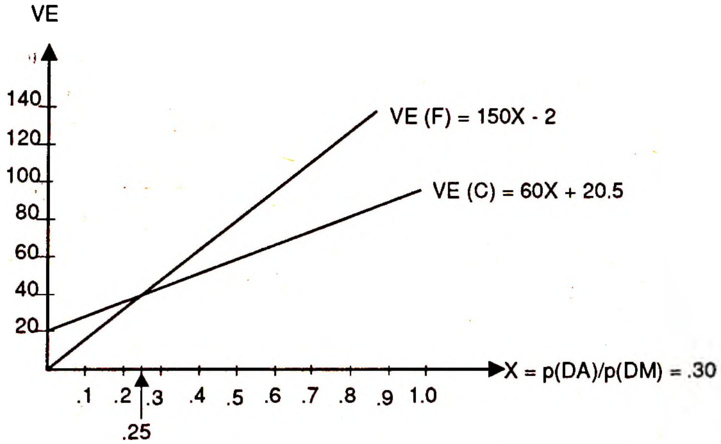
$$VE(F) = 130(X) + 40(0.30) - 20(0.70-X) = 150X - 2 \quad (5)$$

$$VE(C) = 70(X) + 45(0.30) + 10(0.70-X) = 60X + 20.5 \quad (6)$$

En el gráfico 50 se han ilustrado las ecuaciones (5) y (6), que representan los valores esperados de las dos alternativas para diferentes valores de X o p(DA) cuando p(DB) se mantiene en 0.30. Vemos que cuando VE(F) = VE(C), X es igual a 0.25; entonces, para p(DA) = 0.25, p(DM) = 0.30 y p(DB) = 0.45, nos resultará indiferente fabricar o comprar el componente.

Además, el gráfico 50 nos muestra que cuando la probabilidad de DA toma valores mayores que 0.25, VE(F) > VE(C); es decir, que es más conveniente fabricar que comprar el componente. En cambio, cuando la probabilidad de enfrentar una demanda alta es menor que 0.25, las ganancias netas esperadas al fabricar el componente son menores que las ganancias que se obtendrían si se compra, y, por tanto, la decisión óptima variará.

Gráfico 50: Análisis de sensibilidad



Hemos realizado el análisis de sensibilidad respecto a la probabilidad de enfrentar una demanda alta, suponiendo que la probabilidad del nivel medio de demanda se mantiene constante; podríamos efectuar el mismo análisis para cualquiera de las otras dos probabilidades y generar un gráfico similar al anterior.

5. VALOR ESPERADO DE LA INFORMACIÓN PERFECTA (VEIP)

Los expertos de CME afirman que la probabilidad de enfrentar una demanda alta para el equipo de aire acondicionado que fabrica esta compañía es de 0.30, mientras que la probabilidad de una demanda media y baja es de 0.30 y 0.40 respectivamente. Tomando como base esta información se estableció que la alternativa óptima es fabricar el componente requerido. Por otro

lado, el análisis de sensibilidad realizado en la sección anterior nos indica que la decisión de fabricar el componente no será la óptima si la probabilidad de enfrentar una demanda alta es menor que 0.25. Es lógico que la Gerencia de CME quiera mejorar la confiabilidad de sus estimaciones de las probabilidades de la variable de estado mediante la recopilación de información adicional. Esta información puede recogerse, por ejemplo, a través de entrevistas y discusiones con expertos externos, análisis cuantitativo y sistemático de los datos históricos o contratación de un estudio de mercado, y tendrá necesariamente un costo asociado a ella.

Antes de determinar si merece o no la pena recoger esta información adicional, es deseable preguntarse *cuál es el valor de la información perfecta*. Es decir, cuánto estaríamos dispuestos a pagar por deshacernos de la incertidumbre inherente a la variable de estado.

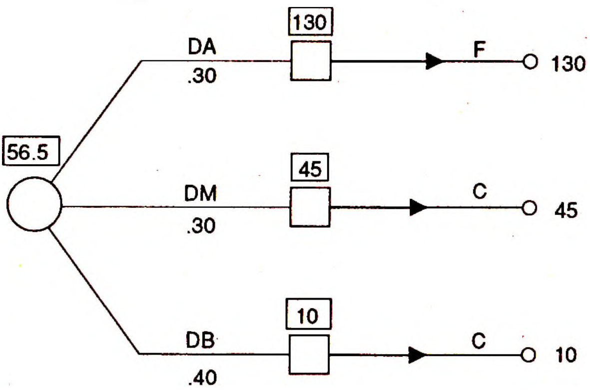
La *información perfecta* es un caso especial de cualquier esquema de recolección de información, y es muy difícil de alcanzar en la práctica. Pero hay dos razones para determinar su valor antes de calcular el valor de cualquier información adicional que sabemos será imperfecta. Primero, el valor de la información perfecta representará el *límite superior* del valor de cualquier información imperfecta. Segundo, su valor sugiere en qué instancia sería conveniente planificar acciones de recolección de información, en el caso que tengamos varias variables de estado en nuestro problema de decisión. Si el valor de la información perfecta es bajo, no será rentable invertir recursos adicionales en conseguir mayor información dado que los costos de obtener esta información podrían superar fácilmente el valor límite establecido por el valor de la información perfecta. En cambio, si el VEIP es alto, podría ser útil dedicar algún esfuerzo a ejecutar programas para mejorar la información.

El valor de la información perfecta es fácil de calcular una vez que la estructura del árbol de decisiones ha sido establecida. El

procedimiento es simplemente cambiar el orden de los nodos en el árbol, colocando el nodo probabilístico que representa la resolución de la incertidumbre antes de cualquier nodo de decisión. El resto del árbol permanece igual. Ilustraremos este procedimiento usando el caso de CME.

Poner el nodo de azar al comienzo del árbol de decisiones significa preguntarnos cuál es el valor esperado de la decisión si podemos elegir el curso de acción después de enterarnos del verdadero estado de la naturaleza; esto es, después de obtener la información perfecta. Así, el gráfico 51 nos muestra el árbol para calcular el valor esperado con información perfecta (VECIP) en el problema de CME.

Gráfico 51: Cálculo del valor esperado de la información perfecta



VECIP = .30 (130) + .30 (45) + .40 (10) = 56.5
VEIP = VECIP - VESIP
= 56.5 - 43 = 13.5

Del gráfico 48 vemos que si CME enfrenta, en efecto, una demanda alta (DA), la mejor alternativa es fabricar el componente, obteniendo una ganancia neta de 130,000 dólares, que es mayor que la ganancia de 70,000 dólares que obtendría si decide comprar. De igual manera, si la demanda es media (DM), la mejor alternativa es comprar obteniendo una ganancia neta de 45,000 dólares; finalmente, si la demanda es baja (DB), la mejor alternativa es comprar obteniendo una ganancia neta de 10,000 dólares. Todo esto está representado en el gráfico 51. La información perfecta le permite a CME, por sí misma, esperar una ganancia neta de 130,000 dólares con una probabilidad de 0.30, una ganancia de 45,000 con una probabilidad de 0.30, y una ganancia de 10,000 con una probabilidad de 0.40. Luego, el VECIP es de:

$$\begin{aligned} VECIP &= 130\,(.30) + 45\,(.30) + 10\,(.40) \\ &= 56.5 \text{ miles de dólares} \end{aligned} \qquad (3)$$

El incremento en el valor esperado debido a la disponibilidad de información perfecta estará dado por:

$$VEIP = VECIP - VESIP \qquad (4)$$

donde VEIP Valor esperado de la información perfecta.
VESIP Valor esperado sin información perfecta.

Del gráfico 49 tenemos que el VESIP es de 43,000 dólares; luego, el VEIP será:

$$\begin{aligned} VEIP &= 56.5 - 43 \\ &= 13.5 \text{ miles de dólares} \end{aligned} \qquad (5)$$

Este valor representa el máximo valor posible en que CME podría ver incrementada su ganancia esperada si cuenta con información cierta con respecto al nivel de demanda de su producto final. Pero sabemos que cualquiera sea el experimento muestral o estudio de mercado a realizarse, este no brindará información «perfecta». En cualquier caso, CME sabe que nunca debe pagar más de 13,500 dólares por cualquier información adicional, no importa cuán buena sea esta. Por ejemplo, si la mejor compañía consultora propone a CME hacer un estudio de mercado para mejorar la evaluación de probabilidades de los diferentes niveles de demanda y solicita 16,000 dólares por realizar dicho estudio, sería absurdo pagar dicha suma por una información imperfecta sabiendo que la información perfecta vale solamente 13,500 dólares.

6. TOMA DE DECISIONES USANDO INFORMACIÓN MUESTRAL

En la sección 5 se estableció la manera de calcular el valor de la información perfecta; pero también se mencionó que es muy difícil, si no imposible, comprar información perfecta. Aun cuando compremos resultados de investigación de mercado de la mejor compañía de mercadotecnia, esta información adicional no será perfecta; es decir, la compañía de mercadotecnia no podrá predecir la demanda de CME con una exactitud absoluta. Si tenemos un historial sobre sus predicciones anteriores podemos establecer las probabilidades sobre la precisión de sus estudios. Estas probabilidades se conocen como la función de verosimilitud, en la terminología del teorema de Bayes. Usaremos este teorema para revisar las probabilidades *a priori* o preliminares sobre la base de la información adicional, bajo el entendimiento de que no lograremos una certidumbre completa.

La búsqueda de información adicional es usualmente lograda a través del diseño de «experimentos» para brindar la información más actualizada acerca de los estados de la naturaleza. Ejemplos de experimentos son la toma de muestras de un embarque para ver la calidad del producto recibido, pruebas médicas e investigación de mercados. Estos experimentos permiten una revisión o actualización de las probabilidades de los estados de la naturaleza. Ilustraremos este proceso usando el problema de CME.

La Compañía de Manufacturas Eléctricas asignó la probabilidad de 0.30 a DA, 0.30 a DM y 0.40 a DB. Usando estas *probabilidades a priori* encontramos que la decisión de fabricar (F) el componente es la óptima, resultando en una ganancia esperada de 43,000 dólares. Además, calculamos que el VEIP es de 13,500 dólares, lo que nos indica que la nueva información acerca del nivel de demanda puede potencialmente valer hasta 13,500 dólares. Ahora supongamos que CME está considerando contratar los servicios de una firma de mercadotecnia para estudiar el tamaño de mercado de su producto. El estudio de investigación de mercado brindará nueva información, la cual puede combinarse con las probabilidades *a priori* a través del teorema de Bayes para obtener probabilidades revisadas o actualizadas de los niveles de demanda, conocidas como *probabilidades posteriores*.

A la información adicional obtenida a través de la «experimentación» se le denota como un *indicador* o *información muestral*, dado que en muchos casos el experimento consiste en tomar una muestra estadística. Usando la terminología de *indicador* (I), denotaremos el resultado del estudio de mercadotecnia de la siguiente manera:

- I Reporte favorable del estudio de mercado; la muestra tomada expresa un interés considerable en el producto de CME.
- I′ Reporte no favorable; la muestra expresa poco interés por el producto de CME.

Dado el resultado del reporte, nuestro objetivo es mejorar los estimados de las probabilidades de los diferentes niveles de demanda tomando como base los resultados del estudio de mercado. Es decir, buscamos encontrar las probabilidades revisadas de la forma p(DA/I), p(DA/I'), p(DM/I), etcétera, donde p(DA/I) representa la probabilidad condicional de que DA ocurra dado que el resultado del estudio de mercado fue favorable (I).

Para usar el teorema de Bayes y poder revisar las probabilidades originales, necesitamos contar con estimados de las probabilidades condicionales de cada uno de los indicadores con respecto a los diferentes niveles de demanda; es decir, necesitamos las *funciones de verosimilitud*: p(I/DA), p(I'/DA), p(I/DM), etcétera. Las fuentes principales para estimar estas funciones de verosimilitud son los datos históricos o estimados subjetivos, que sirven para evaluar la calidad de la predicción de la compañía de mercadotecnia. Así, un estimado de p(I/DA) nos indica con qué frecuencia un estudio de mercado resultó favorable cuando el nivel de demanda fue efectivamente alto.

El registro histórico de la compañía de mercadotecnia en estudios similares ha permitido a la administración de CME estimar las siguientes probabilidades condicionales relevantes:

$$\begin{aligned} p(I/DA) &= 0.6 \\ p(I/DM) &= 0.45 \\ p(I/DB) &= 0.15 \end{aligned} \tag{7}$$

Estos estimados indican que cuando el verdadero nivel de demanda es alto, el estudio de mercado será favorable el 65% de las veces y desfavorable el 35% [p(I'/DA) = 0.35]. Cuando el nivel de demanda es medio, el reporte del estudio de mercado será favorable el 45% de las veces, mientras que si el nivel de demanda es bajo, el reporte sólo será favorable el 15% de las veces. Ahora veamos cómo podemos usar esta información adicional en el proceso de toma de decisiones.

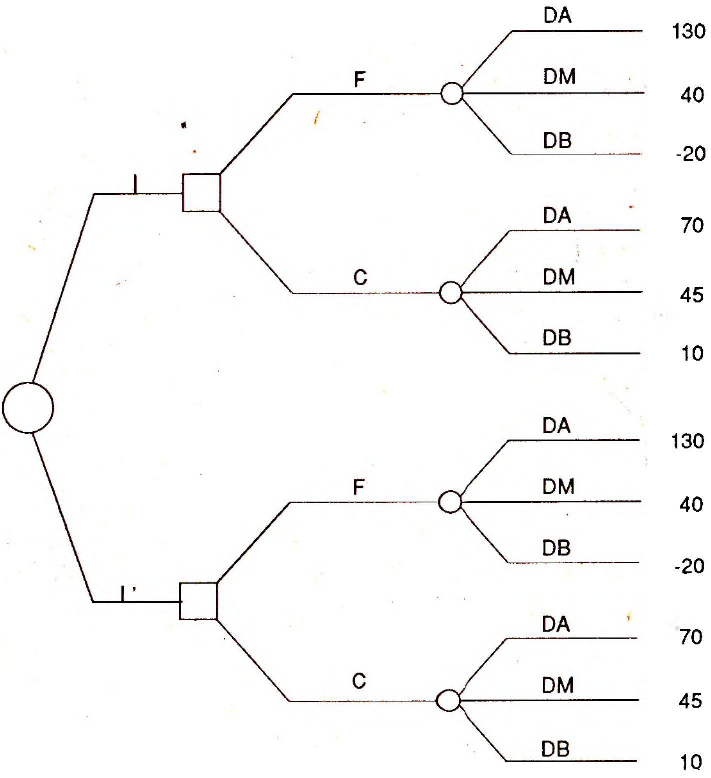
El gráfico 52 muestra el árbol de decisiones relevantes si se decide encargar el estudio de mercado. A medida que nos trasladamos de izquierda a derecha, el árbol nos muestra el orden cronológico del proceso de toma de decisiones. En primer lugar, CME obtendrá el reporte del estudio de mercado (I o I'); luego, sobre la base de este reporte, CME tomará una decisión (F o C); y finalmente, se verificará el tamaño de mercado específico que en efecto se enfrentará (DA, DM o DB). La combinación de la decisión y el nivel de demanda definirá el nivel de ganancias netas, las cuales se muestran al final del árbol de decisiones.

Note que en el gráfico 52 hemos usado un nodo de azar con dos ramas, I e I', para denotar los resultados del estudio de mercado porque estos resultados no están bajo el control del decisor sino que están determinados por el azar. Al final de estas ramas tenemos un nodo de decisión en el que el decisor debe elegir entre fabricar o comprar el componente. Seleccionar la mejor rama de decisión es equivalente a tomar la mejor decisión. Sin embargo, para seleccionar la mejor rama necesita conocer el valor esperado en cada una de estas ramas. Para calcular estos valores es necesario conocer las probabilidades asociadas con cada una de las ramas DA, DM y DB que están al final de cada rama de decisión. Recordemos que las ramas de azar están fuera del control del decisor, y que dependen de la probabilidad asociada con cada una. Entonces, antes de continuar con el análisis del árbol de decisiones y establecer una estrategia de decisión, debemos completar la información en el árbol de decisiones del gráfico 52 calculando las probabilidades asociadas con cada rama de azar tanto para el resultado del estudio de mercado como para el nivel de demanda. Dado que los niveles de demanda se verifican después de que los resultados del estudio de mercado han sido revelados, necesitamos calcular las probabilidades condicionales o probabilidades posteriores: p(DA/I), p(DA/I'), p(DM/I), etcétera.

Usando el concepto de expansión en cadena explicado en el capítulo II podemos escribir las probabilidades conjuntas de la siguiente manera:

$p\,(DA,I)$	=	$p(DA)$	*	$p(I/DA)$
$p\,(DA,I')$	=	$p(DA)$	*	$p(I'/DA)$
$p\,(DB,I')$	=	$p\,(DB)$	*	$p(I'/DB)$
		Probabilidades previas		Función de verosimilitud

Gráfico 52: Compañía de Manufacturas Eléctricas: Árbol de decisiones con estudio del mercado



Estas probabilidades conjuntas pueden calcularse, dado que conocemos tanto las probabilidades previas (gráfico 48) como las funciones de verosimilitud. Así, por ejemplo, la probabilidad conjunta de obtener un nivel de demanda alto y de que el informe sea favorable es:

$$p(DA,I) = p(DA) * p(I/DA)$$
$$= (0.30)\ (0.65) = 0.195 \qquad (9)$$

Usando una vez más el concepto de expansión en cadena, esta probabilidad conjunta también puede escribirse de la siguiente manera:

$$p(DA,I) = p(I)\ p(DA/I) \qquad (10)$$

De donde puede despejarse la probabilidad posterior, p(DA/I), obteniendo:

$$p(DA/I) = \frac{p(DA,I)}{p(I)} \qquad (11)$$

El valor de p(I), la probabilidad de obtener un reporte favorable del estudio de mercado, conocida como la *probabilidad pre-posterior,* se obtiene sumando las probabilidades conjuntas:

$$p(I) = p(DA,I) + p(DM,I) + p(DB,I) \qquad (12)$$

Estas probabilidades conjuntas se calcularon multiplicando las probabilidades previas por la función de verosimilitud, como se muestra en el *árbol de asignación de probabilidades* del gráfico 53. Reemplazando estos valores en la ecuación (12), obtenemos:

$$p(I) \quad = 0.195 + 0.135 + 0.060 = 0.39 \qquad (13)$$

Luego, la probabilidad posterior, expresada en la ecuación (11), será:

$$p(DA/I) = \frac{0.195}{0.390} = 0.50 \qquad (14)$$

Este es el procedimiento bayesiano para calcular las probabilidades revisadas, las cuales se presentan en el *árbol cronológico* del gráfico 53. Como su nombre lo indica, este árbol representa las variables de estado en el orden cronólogico en que suceden; primero se conoce el resultado del estudio de mercado, luego el nivel de demanda que enfrentará CME por su producto. Resumiendo, podemos decir que hay tres pasos para calcular las probabilidades posteriores o revisadas. Primero calculamos las probabilidades conjuntas en el árbol de asignación de probabilidades y las trasladamos a los nodos finales del árbol cronológico. Segundo, calculamos las probabilidades pre-posteriores, sumando las probabilidades conjuntas respectivas. Finalmente, calculamos las probabilidades revisadas usando la relación de probabilidad condicional (ecuación 11). Debemos notar que en el árbol cronológico también se cumple la relación de expansión en cadena, la que implica que las probabilidades en los nodos finales son iguales al producto de las probabilidades de las ramas que van desde el nodo inicial hasta dicho nodo final. Así:

$$\begin{aligned} p\,(I'\,, DB) &= p\,(I')\,p(DB/I') \\ &= (0.61)\,(0.56) = 0.34 \end{aligned}$$

Las probabilidades posteriores representan las probabilidades estimadas de cada nivel de demanda dado el resultado del

estudio de mercado; p(DA/I) = 0.50, por ejemplo, indica que hay una probabilidad de 0.50 de que el nivel de demanda sea alto cuando se obtiene un reporte favorable en el estudio de mercado. Sin embargo, vemos que el nivel de demanda del producto de CME puede ser medio o bajo, a pesar de que el reporte sea favorable. Así, p(DM/I) = 0.35 y p(DB/I) = 0.15. Pero debemos notar que las probabilidades originales (previas) han sido revisadas dado el reporte favorable; así, la probabilidad de DA pasó de 0.30 a 0.50 [(p(DA) = 0.30; ahora, p(DA/I) = 0.50)]. De igual manera, las probabilidades revisadas dado un reporte desfavorable han variado con respecto a las probabilidades previas [(*e.g.*, p(DA) = 0.30; ahora, p(DA/I') = 0.17)] (ver el árbol cronológico del gráfico 53).

Con las probabilidades de los resultados del estudio de mercado, p(I) y p(I'), y las probabilidades posteriores podemos completar nuestro árbol de decisiones con toda la información requerida para su solución, como se muestra en el gráfico 54. Estamos listos para desarrollar una estrategia óptima de decisión para CME si es que el estudio de mercado se lleva a cabo.

7. ESTRATEGIA ÓPTIMA DE DECISIÓN CON INFORMACIÓN MUESTRAL

Con el árbol de decisiones completo, ahora podemos resolver nuestro problema para encontrar la decisión óptima para CME. Como se señaló en la sección 3, para resolver el árbol de decisiones nos trasladamos conceptualmente a los nodos finales del árbol y trabajamos hacia atrás utilizando dos mecanismos: calculamos el valor esperado en cada nodo de azar, y elegimos la mejor alternativa en cada nodo de decisión. Así, en el gráfico 54 primero calculamos los valores esperados para los nodos de azar que representan el nivel de demanda. Para el nodo de azar que está después de las ramas I y F, por ejemplo, tenemos:

$$VE(I,F) = 0.50(130) + 0.35(40) + 0.15(-20) = 76$$

Gráfico 53: Cálculo de probabilidades posteriores o revisadas

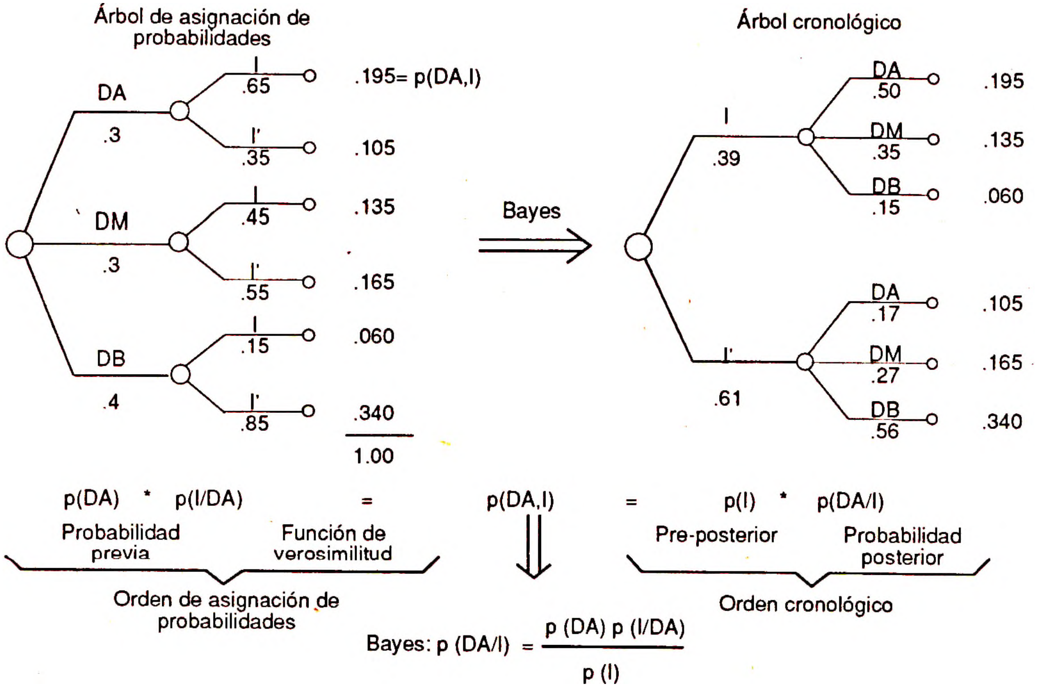
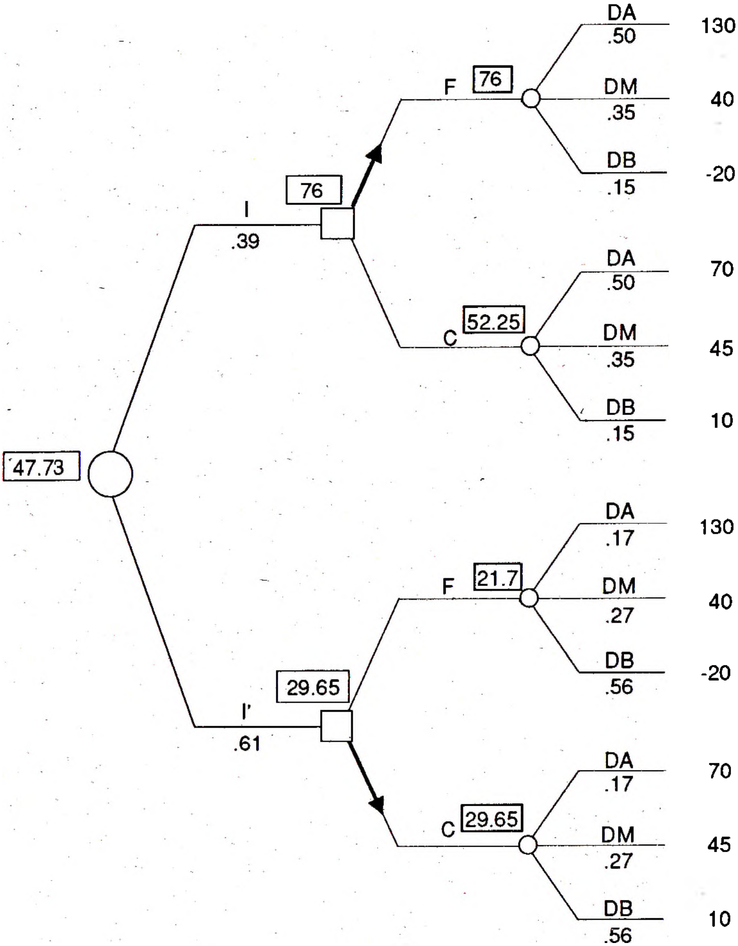


Gráfico 54: Solución del árbol de decisiones del problema de CME, con la realización de estudios de mercados



Y diremos que el VE de tener un informe favorable (I) y de fabricar el componente (F) es de 76,000 dólares. Este valor está escrito encima del nodo de azar respectivo. De igual manera se calcularon los valores esperados para los otros nodos de azar:

$$VE(I,C) = 0.50(70) + 0.35(45) + 0.15(10) = 52.25$$

$$VE(I',F) = 0.17(130) + 0.27(40) + 0.56(-20) = 21.70$$

$$VE(I',C) = 0.17(70) + 27(45) + 0.56(10) = 29.65$$

Continuamos trabajando hacia atrás en el árbol de decisiones y nos encontramos en los nodos de decisión. Aquí el decisor debe elegir F o C, dado el resultado del reporte del estudio de mercado. Como el decisor está tratando de maximizar las ganancias netas esperadas, su elección es clara: si el reporte es favorable (I), la decisión óptima es fabricar el componente (F); si el reporte no es favorable (I'), la decisión óptima es comprar el componente (C). Es decir, si el reporte es favorable y dado que la decisión óptima de fabricar el componente resulta en un valor esperado de 76,000 dólares, diremos que el VE en el nodo de decisión es de 76,000 dólares, y anotamos este valor encima de dicho nodo. De manera similar, si el reporte no es favorable y dado que la decisión óptima de comprar (C) ofrece un valor esperado de 29,650 dólares, anotamos este valor encima del nodo de decisión. Resumiendo, para los nodos de decisión que están después de las ramas del resultado del estudio de mercado, tenemos:

$$VE(I) = 76{,}000$$

$$VE(I') = 29{,}650$$

Como último paso, continuamos trabajando hacia atrás al nodo de azar que representa al indicador de mercado y calculamos su valor esperado:

$$\begin{aligned} VE\text{ (con estudio de mercado)} &= 0.39\,(76{,}000) + 0.61\,(29{,}650) \\ &= 47{,}730 \text{ dólares} \end{aligned}$$

Los 47,730 dólares representan el valor esperado de la estrategia óptima de decisión cuando se encarga realizar el estudio de mercado. También se le conoce como el VE de usar información muestral.

Debemos señalar que la decisión final todavía no se ha determinado. Necesitamos conocer el resultado del estudio de mercado antes de decidir si fabricar (F) o comprar (C) el componente para la fabricación de los aparatos de aire acondicionado. Sin embargo, los resultados del análisis de la teoría de decisiones nos definen la siguiente estrategia de decisión óptima, si el estudio de mercado se lleva a cabo:

- Si el reporte es favorable (I), entonces se debe fabricar el componente, para maximizar las ganancias esperadas.
- Si el reporte no es favorable (I'), entonces se debe comprar el componente.

Hemos visto cómo los árboles de decisiones pueden usarse para desarrollar estrategias óptimas en problemas de decisión en un mundo incierto cuando es posible realizar "experimentos" para conseguir información adicional. Si bien otros problemas reales de decisión pueden ser más complejos que el problema de CME, el enfoque que hemos presentado es igualmente aplicable. En primer lugar, el analista debe trazar el árbol de decisiones consistente de nodos y ramas que representan el resultado del experimento; luego se grafican las decisiones alternativas y las variables de azar de tal manera que el árbol describa el proceso de toma de decisiones específicas. Debemos señalar que la diferencia entre un árbol de decisión sin información adicional y uno con información experimental es que este último tiene un nodo de azar adicional al comienzo del árbol, que representa el resultado de la recolección de información (indicador). El nodo de azar que representa la variable de estado en cuestión (*e.g.*, nivel de demanda) permanece en su lugar original, pero las ramas que salen de este nodo tienen nuevas probabilidades que se asignan

sobre la base de la información adicional recibida. En el árbol original se usaron las probabilidades previas, pero en el árbol con información adicional se usan las probabilidades revisadas o posteriores. Es decir que será necesario aplicar el teorema de Bayes para determinar las probabilidades del resultado de la información adicional y las probabilidades revisadas de la variable de estado. Luego, trabajando hacia atrás en el árbol, calculando valores esperados en los nodos de azar y eligiendo la mejor rama de decisión en los nodos de decisión, el analista puede determinar la estrategia óptima de decisión y el valor esperado asociado al problema.

8. VALOR ESPERADO DE LA INFORMACIÓN MUESTRAL

Al igual que calculamos el valor esperado de la información perfecta (VEIP) en la sección 4, aquí podemos calcular el valor esperado de la información muestral (VEIM). En el problema de CME establecimos una estrategia óptima de decisión de fabricar el componente si el reporte del estudio de mercado es favorable y de comprar el componente si el reporte no es favorable. Pero dado que la información adicional suministrada por la compañía de mercadotecnia resultará en un costo adicional para CME, será importante determinar el valor de esta nueva información. Este valor esperado incremental de la información muestral estará dado por:

$$VEIM = VECIM - VESIM$$

donde VECIM Valor esperado con información muestral.
VESIM Valor esperado sin información muestral.

Del gráfico 54 tenemos que el VECIM, en el problema de CME, es de 47,730 dólares; y del gráfico 49 el VESIM es de 43,000 dólares. Entonces el VEIM será:

$$VEIM = 47{,}730 - 43{,}000 = 4{,}730 \text{ dólares}$$

Luego, CME debería estar dispuesta a pagar hasta 4,730 dólares por la información del estudio de mercado.

En la sección 4 vimos que el VEIP para el problema de CME fue de 13,500 dólares. Si bien no podemos esperar que el estudio de mercado nos dé información perfecta, podemos usar una medida de *eficiencia* para expresar el valor relativo del reporte. Considerando que la información perfecta tiene una eficiencia del 100%, la eficiencia (E) de la información muestral estará dada por:

$$E = \frac{VEIM}{VEIP} * 100$$

En el problema de CME, tenemos:

$$E = \frac{4{,}730}{13{,}500} * 100 = 35\%$$

Esto quiere decir que la información del estudio de mercado es sólo el 35% de eficiente en relación a la información perfecta. Una eficiencia baja nos hará buscar otros tipos de información muestral, como por ejemplo otra compañía de mercadotecnia. En cambio, una alta eficiencia (E) nos indicará que no será necesario buscar fuentes de información adicional.

● *Ejercicios*

1. ¿En qué difieren los métodos para la toma de decisiones de este capítulo con los estudiados en el capítulo V?

2. Una empresa electrónica está tratando de decidir si debe producir o no un nuevo sistema de *beepers*. La decisión de producirlo significa una inversión de 5 millones de dólares y no se conoce la demanda de este instrumento. Si la demanda es alta, la empresa espera un rendi-

miento de 11 millones de dólares. Si la demanda es moderada, el rendimiento será de 7.4 millones de dólares; y si la demanda es baja significaría un rendimiento de sólo 4.2 millones de dólares, que no alcanzaría a cubrir los gastos de inversión. Se estima que la probabilidad de obtener una demanda baja es de 0.1, mientras que la probabilidad de una demanda alta es de 0.5.

a. Si la compañía se basa en el criterio del valor esperado, ¿debería efectuar la inversión?

b. Un miembro del Directorio propuso que se tome una muestra para estimar la demanda real. ¿Cuál es el monto máximo que se debe pagar por este estudio?

c. ¿Entre qué límites puede variar la probabilidad de que la demanda sea alta sin cambiar la decisión elegida en (a)?

3. El Gerente de una planta manufacturera que produce un artículo para el cual la demanda es muy variada, debe decidir entre comprar una nueva máquina para cierta línea de montaje o reparar la máquina que actualmente presta servicio en esa línea. El costo de la compostura es de 500 dólares, mientras que el costo de la nueva máquina es de 7,000 dólares. Con la máquina nueva, el costo variable de producción por artículo es $ 0.60; con la máquina reparada el costo variable unitario es $ 1.10. El artículo se produce a medida que es demandado, para no quedar con excedentes. El precio de venta de cada unidad del artículo es de 2 dólares y el Gerente estima que la demanda ocurrirá de acuerdo con la siguiente distribución:

Demanda	5,000	10,000	25,000	50,000
Probabilidad	0.2	0.4	0.3	0.1

a. Si el objetivo del Gerente es maximizar las ganancias esperadas de la planta, ¿cuál es la decisión óptima?

b. ¿Cuál es el valor esperado de la información perfecta respecto al nivel de demanda?

c. Suponga que el Gerente tiene además la alternativa de comprar otra máquina, que cuesta 2,000 y tiene un costo de operación por unidad de 1 dólar. ¿Cuál es ahora la decisión óptima?

4. Una agencia de publicidad ha diseñado tres campañas alternativas para uno de sus clientes. El cliente estudió las propuestas y construyó una distribución de probabilidades que describe el valor presente de las ganancias netas esperadas si invierte en una campaña de publicidad específica. Las distribuciones de probabilidades de los tres programas propuestos son las siguientes:

Campaña de publicidad	Valor presente neto de las ganancias			
	-30,000	-10,000	-50,000	100,000
A	0.10	0.30	0.40	0.20
B	0.15	0.25	0.50	0.10
C	0.00	0.30	0.65	0.05

a. Grafique el árbol de decisiones que describe las opciones de decisión.
b. Usando el criterio del valor esperado, ¿cuál es la decisión óptima?
c. ¿Cuál es la probabilidad de que el valor presente neto de las ganancias sea menor que cero en cada una de las campañas de publicidad? ¿Cuál de los programas propuestos adoptaría usted? Explique su elección.

5. Carlos Rodríguez está programando un espectáculo de pelea de toros el 18 de agosto, con motivo del aniversario de la fundación de la ciudad de Arequipa. Las ganancias que se obtengan dependerán en gran medida del clima en el día del evento. En concreto, si el día es lluvioso, Carlos pierde 15,000 dólares; si es soleado, gana 10,000 dólares. Se supone que los días o son lluviosos o soleados. Carlos puede decidir cancelar el evento, pero si lo hace pierde el depósito de 1,000 dólares efectuado al alquilar el local. Los registros del pasado indican que en la quinta parte de los últimos cincuenta años, ha llovido en esa fecha.
a. Establezca la tabla de resultados y grafique el árbol de decisiones.
b. ¿Qué decisión debe tomar Carlos para maximizar su beneficio neto esperado en dólares?
c. ¿Entre qué límites puede variar la probabilidad de que llueva, sin que cambie la decisión establecida en (b)?
d. ¿Cuál es el monto máximo que Carlos estará dispuesto a pagar por conocer de antemano el clima del 18 de agosto?
e. La "Brujita Coty", una clarividente muy famosa, ofrece sus servicios a Carlos. En las ocasiones que ha llovido, Coty acertó el 90% de las veces. Por otra parte, cuando predijo un día soleado, acertó sólo el 80% de las veces. ¿Cuánto estaría dispuesto a pagar Carlos por los servicios de Coty?
f. Si Carlos decide contratar a la clarividente, ¿qué estrategia debe seguir para maximizar su beneficio neto esperado?

6. "Tortas Delicia" planea su producción diaria decidiendo producir 100, 200 ó 300 tortas. El cuadro de ingresos netos asociados a diferentes niveles de demanda y producción es el siguiente:

	Demanda (tortas)		
Producción	100 D1	200 D2	300 D3
100 P1	500	200	-100
200 P2	-400	800	700
300 P3	-1,000	-200	1,600

a. Si p(D1) = 0.30 y p(D2) = 0.20, ¿cuál sería el nivel óptimo de producción?

b. ¿Cuál es el valor máximo que "Tortas Delicia" debería pagar por conocer de antemano el nivel de demanda?

c. Algunos días "Tortas Delicia" recibe pedidos por teléfono. Denotemos:

I1 Se reciben pedidos por teléfono.

I2 No se reciben pedidos por teléfono.

Si se sabe que: p(I2/D1) = 0.80, p(I2/D2) = 0.40 y p(I2/D3) = 0.10, ¿cuál es la producción óptima en los días que "Tortas Delicia" no recibe pedidos por teléfono?

7. Un inversionista ganará 12,000 dólares en intereses si deja sus fondos en una cuenta de ahorros, pero está considerando retirarlos para invertirlos en acciones de la compañía "Vida Feliz", que tiene una probabilidad de 70% de rendir utilidades de 20,000 dólares y una probabilidad de 30% de rendir utilidades de 5,000 dólares, si las acciones son compradas hoy. Si el inversionista espera, hay una probabilidad del 70% de que las acciones empiecen a subir. Si esto pasa, el inversionista comprará acciones y las probabilidades de rendimiento serán de 99 a 1 de rendir 18,000 ó 3,000 dólares. Si el precio de las acciones no sube, el inversionista dejará sus fondos en el banco.

a. ¿Cuál es la mejor alternativa, usando el criterio del valor esperado?

b. ¿Cuál es el valor esperado de la información perfecta con respecto a si las acciones subirán, en el caso de que el inversionista decida esperar?

8. La compañía inmobiliaria "Condominios del Cielo" ha comprado recientemente un terreno en La Molina y está tratando de determinar el tamaño del complejo de condominios que debe construir. Se están considerando tres tamaños de acuerdo con el número de casas que conforme el complejo: pequeño (P), mediano (M) y grande (G). Al mismo tiempo, la incertidumbre en la evolución de la economía nacio-

nal dificulta predecir la demanda por nuevos condominios. El Gerente de "Condominios del Cielo" sabe que si se construye un complejo grande y se enfrenta una demanda baja, podría ser muy costoso para la compañía. Asimismo, si se construye un complejo pequeño y la demanda es alta, las utilidades de la firma serán mucho menores que si se hubiese construido un complejo mayor. Con los tres niveles de demanda (baja, intermedia y alta), el Gerente ha preparado el siguiente cuadro de utilidades netas (en miles de dólares):

	Demanda		
Decisión	Baja(B)	Intermedia (I)	Alta (A)
Pequeño (P)	400	400	400
Mediano (M)	100	600	600
Grande (G)	-300	300	900

a. Si p(B) = 0.20, p(I) = 0.35, ¿cuál es la decisión que usted recomendaría, usando el criterio del valor esperado?
b. ¿Cuál es el valor esperado de la información perfecta?
c. La compañía está considerando llevar a cabo un estudio de mercado, el cual ayudará a evaluar la demanda para el nuevo complejo de condominios. El estudio arrojará un resultado I, que consiste en tres indicadores de demanda: débil (D), típica (T) y fuerte(F), con las siguientes funciones de verosimilitud:

Demanda	Indicadores de demanda		
	D	T	F
B	0.6	0.3	0.1
I	0.4	0.4	0.2
A	0.1	0.4	0.5

¿Cuál es el valor esperado del estudio de mercado?

9. Un inventor ha desarrollado un instrumento electrónico para controlar las impurezas en el metal producido en un proceso de fundi-

ción. Una compañía que vende este tipo de equipo está considerando la compra de los derechos de patente de este instrumento por 400,000 dólares. La compañía cree que hay una posibilidad entre cinco de que el instrumento sea un éxito de ventas. Se estima que si el instrumento tiene éxito producirá ingresos netos de 4.5 millones de dólares. Si el producto no tiene éxito no se recibirán ingresos.

a. Dibujar un árbol de decisión que describa las opciones de decisión y determine la mejor política.

b. Si hay un grupo de investigación de mercados que puede ofrecer información perfecta respecto al éxito de este producto, ¿cuál será la cantidad máxima que la compañía estará dispuesta a pagar por este servicio?

c. Suponga que el grupo de investigación de mercados puede hacer un sondeo del mercado que con probabilidad de 0.70 dará un resultado favorable si el instrumento es un éxito y con probabilidad 0.9 dará un resultado desfavorable si el instrumento no se vende. ¿Cuánto valdrá para la compañía este estudio?

10. Una compañía manufacturera necesita decidir si debe o no emprender una campaña de publicidad para un producto cuyas ventas han estado estancadas. El costo de la campaña es de 100,000 dólares. Se estima que una campaña muy exitosa incrementaría las utilidades en 400,000 dólares (de las cuales deberán sustraerse los costos de la campaña). Una campaña moderadamente exitosa incrementaría las utilidades en 100,000 dólares; mientras que una campaña desfavorable no incrementaría en nada las utilidades. De los registros históricos, el 50% de campañas similares han sido muy exitosas, el 30% moderadamente exitosas y el resto nada exitosas.

La compañía tiene la oportunidad de contratar los servicios de una instalación de mercadotecnia para evaluar la efectividad potencial de la campaña publicitaria. El historial de esta compañía de mercadotecnia es tal que ha reportado favorablemente en el 80% de las campañas que resultaron altamente favorables, el 40% de aquellas que fueron moderadamente favorables y 10% de las campañas desfavorables.

a. Sin contratar los servicios de mercadotecnia, ¿debería emprenderse la campaña publicitaria, usando el criterio del valor esperado?

b. ¿Cuál es el valor esperado de la información perfecta?

c. Encuentre las probabilidades posteriores o revisadas para los tres estados de la naturaleza, dado el reporte de la compañía de mercadotecnia.

d. Establezca una estrategia de decisión, dado el reporte de la compañía de mercadotecnia.

e. ¿Cuál es el valor esperado del estudio de la compañía de mercadotecnia?

11. Una editorial está considerando lanzar una nueva revista mensual con artículos e información para inversionistas. Basado en su experiencia y en sus percepciones, el Gerente de la editorial ha establecido una tabla de ganancias anuales considerando tres niveles distintos de demanda por su revista.

	Ganancias anuales (En dólares)	
Demanda	Lanzar la revista	No lanzarla
Baja (DB)	-250,000	0
Regular (DR)	50,000	0
Alta (DA)	300,000	0

El Gerente estima además que las probabilidades de estos niveles de demanda son: p(DB) = 0.5, p(DR) = 0.2 y p(DA) = 0.3.

a. Sobre la base de esta información y usando el criterio del valor esperado, ¿debe lanzarse la nueva revista?

b. ¿Cuál es el valor esperado de la información perfecta?

El Gerente de la editorial tiene la oportunidad de hacer una prueba para determinar la aceptación que tendrá la revista y basar su decisión en los resultados de esta prueba. La prueba arroja un diagnóstico favorable (F) y uno no favorable (F'). Basado en experiencias previas en relación a otras publicaciones, el Gerente ha establecido las siguientes probabilidades condicionales dadas las posibles demandas:

$$p(F/DB) = 0.1 \quad p(F/DR) = 0.6 \quad p(F/DA) = 0.7$$

c. ¿Cuál es la probabilidad de que el diagnóstico sea favorable?

d. ¿Cuál es la mejor decisión para la editorial si la prueba resulta en un diagnóstico favorable?

e. ¿Cuál es la mejor decisión si el resultado es desfavorable?

f. ¿Cuál es el valor esperado de la prueba de aceptación de la revista?

12. El Gerente de Investigación de "Petroéxito" está tratando de decidir si debe asignar fondos a un proyecto para producir un nuevo lubricante. Se piensa que el proyecto puede ser un gran éxito técnico, uno menor o un fracaso completo. La compañía ha estimado que el

valor de un gran éxito técnico es de 150,000 dólares, dado que el lubricante podrá ser utilizado en un gran número de los productos que elabora. Si el proyecto es un éxito menor, su valor es de 10,000 dólares, dado que "Petroéxito" piensa que el conocimiento ganado será útil para futuros proyectos. Si el proyecto es un fracaso, le costará 100,000 dólares a la compañía.

Basándose en la opinión de los científicos involucrados y en la propia apreciación del Gerente, las probabilidades a *priori* son:

p(gran éxito) = 0.15
p(fracaso) = 0.40

a. Usando el criterio de valor esperado, ¿debería llevarse a cabo el proyecto?
b. ¿Cuál es el valor esperado de la información perfecta sobre el éxito técnico que se logrará con el nuevo lubricante?

Existe un experimento que puede llevarse a cabo para saber más sobre la factibilidad técnica del proyecto. Este experimento tiene tres posibles resultados:

I1: El lubricante trabaja bien a cualquier temperatura.
I2: El lubricante prototipo sólo trabaja bien a temperaturas mayores de 10° C.
I3: El lubricante prototipo no trabaja bien a cualquier temperatura.

Suponga que podemos determinar las siguientes probabilidades condicionales:

p(I1/gran éxito)	= 0.70	p(I2/gran éxito)	= 0.25
p(I1/menor éxito)	= 0.10	p(I2/menor éxito)	= 0.70
p(I1/fracaso)	= 0.10	p(I2/fracaso)	= 0.30

c. Suponga que el experimento se lleva a cabo, y se encuentra que el lubricante prototipo trabaja bien a cualquier temperatura. ¿Debería llevarse a cabo el proyecto?
d. Suponga que el experimento se lleva a cabo y se encuentra que el prototipo sólo funciona bien a temperaturas mayores a 10° C. ¿Debería llevarse a cabo el proyecto?
e. ¿Cuál es el valor esperado del experimento?

13. Una cervecería de la capital está considerando la introducción de una nueva bebida no alcohólica. La compañía piensa que hay una probabilidad del 60% de que el producto sea un éxito. El cuadro de ingresos netos es el siguiente:

	Éxito (E)	Fracaso (E′)
Producir	$ 250,000	$ -300,000
No producir	$ -50,000	$ -20,000

La cervecería tiene la posibilidad de contratar a dos compañías de marketing para obtener información adicional. La compañía "Mauro Marketing" afirma que ha desarrollado un indicador de mercado, I, para el cual se tiene $p(I/E) = 0.7$ y $p(I/E') = 0.4$. La compañía "Nuevo Mundo" afirma que ha desarrollado un indicador de mercado, J, para el cual $p(J/E) = 0.6$ y $p(J/E') = 0.3$.

a. Usando el criterio del valor esperado, sin contratar a una compañía de marketing, ¿cuál es la decisión óptima para la cervecería?

b. ¿Cuál es el valor esperado de la información perfecta?

c. Encuentre el valor esperado de la información de "Mauro" y el de "Nuevo Mundo".

d. Si ambas firmas cobran 5,000 dólares, ¿a qué compañía debe contratar la cervecería? ¿Por qué?

14. Un procedimiento de control de calidad implica la inspección del 100% de las piezas recibidas de un abastecedor. Los registros históricos muestran las siguientes tasas de piezas defectuosas.

Porcentaje de piezas defectuosas	Probabilidad
0	0.15
1	0.25
2	0.40
3	0.20

El costo para el control de calidad con una inspección del 100% es de 250 dólares para cada embarque de 500 piezas. Si no se hace esta

inspección del 100%, las piezas se instalan sin inspección previa, y las partes defectuosas se cambian después, pero implican un costo adicional de 25 dólares por cada pieza defectuosa.
a. Complete el siguiente cuadro de costos, para cada una de las alternativas:

	Porcentaje de piezas defectuosas			
	0%	1%	2%	3%
Inspección	250	250	250	250
Sin inspección				

b. El Gerente está considerando eliminar el proceso de inspección para ahorrar 250 dólares por embarque. ¿Recomendaría usted esta acción?
c. ¿Cuál es el monto máximo que el Gerente estará dispuesto a pagar por conocer de antemano la tasa de piezas defectuosas?

15. Un consultor está considerando presentar propuestas detalladas para dos posibles contratos. La preparación de la propuesta para el primer contrato cuesta 1,000 dólares, mientras que la preparación de la propuesta para el segundo cuesta 1,500 dólares . Si la propuesta para el primero es aceptada y se realiza el trabajo, se obtendría una ganancia de 8,000 dólares. Si la propuesta para el segundo es aceptada y se realiza el trabajo, se logrará una ganancia de 12,000 dólares. Los costos de preparación de las propuestas deben ser deducidos de estas ganancias. El consultor puede presentar propuestas para ambos contratos, si lo desea. Pero no tiene los recursos para realizar ambos trabajos simultáneamente. Si una propuesta se presenta y es aceptada, y el consultor no puede realizar el trabajo, él considera esto como un costo por pérdida de prestigio y lo evalúa como un costo de 2,000 dólares.
a. El consultor tiene cuatro alternativas de decisión. ¿Cuáles son?
b. El consultor cree que existe una probabilidad de 0.80 de que la propuesta para el primer contrato sea aceptada, y una probabilidad de 0.50 de que la segunda propuesta sea aceptada. Además, piensa que la aceptación de una propuesta es independiente de la aceptación de la otra. ¿Cuál es el árbol de decisiones para el problema del consultor?

c. Tomando como base el criterio del valor esperado, ¿cuál es la alternativa que debe escoger el consultor, y cuál es el valor esperado de sus ganancias?
d. ¿Cuánto estará dispuesto a pagar el consultor por conocer de antemano si ganará o no el primer contrato?
e. ¿Cuánto estará dispuesto a pagar el consultor para conocer de antemano si ganará o no el segundo contrato?

16. El restaurante "El Sombrero" está considerando abrir una nueva sucursal en la avenida Larco de Miraflores. Tiene tres modelos diferentes, con diferentes capacidades. "El Sombrero" estima que el número promedio de clientes por hora será de 80, 100 ó 120. El cuadro de ganancias netas por semana para los tres modelos es el siguiente:

	Promedio de clientes por hora		
	80	100	120
Modelo A	$ 10,000	15,000	14,000
Modelo B	8,000	18,000	12,000
Modelo C	6,000	16,000	21,000

"El Sombrero" estima que la probabilidad de tener 80 clientes por hora es la misma que la probabilidad de tener 120 clientes por hora y el doble de la probabilidad de tener 100 clientes por hora.
a. Usando el criterio del valor esperado, ¿cuál es la decisión óptima?
b. ¿Cuál es el valor de la información perfecta?
c. "El Sombrero" puede contratar un estudio de mercado por 1,000 dólares. Los resultados del estudio son "favorables" o "desfavorables". Las probabilidades condicionales son:

p(favorable/80 clientes por hora) = 0.2
p(favorable/100 clientes por hora) = 0.5
p(favorable/120 clientes por hora) = 0.9

¿Debería "El Sombrero" contratar el estudio?

17. Una compañía produce un alimento perecedero a un costo de 10 dólares por caja, siendo su precio de venta de 15 dólares por caja. Con el objeto de planificar la producción, la compañía considera tres niveles

de demanda diaria: 100, 200 ó 300 cajas. Si la demanda es menor que lo producido, el exceso de producción se pierde. Si la demanda es mayor que la producción, la compañía, en un intento de conservar la buena imagen de servicio, satisface el exceso de demanda con una producción especial a un costo de 18 dólares por caja. Sin embargo, el producto siempre se vende a 15 dólares.

a. Si p(100) = 0.2, p(200) = 0.2 y p(300) = 0.6, use el criterio del valor esperado para determinar el nivel de producción diaria.

b. ¿Cuál es el valor esperado de la información perfecta?

c. Una consultora de mercadotecnia ofrece hacer un estudio de mercado para predecir la demanda, por el cual cobra 650 dólares. ¿Debería contratarse sus servicios?

18. La tienda de departamentos "Diamante" acaba de adquirir la cadena de tiendas "Chicho e Hijos". "Diamante" ha recibido una oferta de "Oshborn" para comprar la tienda "Chicho" en la avenida Primavera por 120,000 dólares. "Diamante" ha determinado que la utilidad neta potencial para esta tienda dependerá del estado de la economía nacional. Los estimados de las utilidades netas de la tienda son 80, 100, 120 ó 140 miles de soles con las siguientes probabilidades:

p(80) = 0.20
p(100) = 0.30
p(120) = 0.10
p(140) = 0.40

a. Sobre la base de esta información, ¿debería "Diamante" vender la tienda?

b. ¿Cuál es el monto máximo que "Diamante" estará dispuesto a pagar por conocer de antemano el estado de la economía nacional?

c. "Diamante" podría tener una proyección del comportamiento de la economía a un costo de 10 mil soles. La proyección económica produce un indicador, I, para el cual se tiene:

p(I/80) = 0.1 p(I/100) = 0.2
p(I/120) = 0.6 p(I/140) = 0.8

¿Debe la compañía hacer la proyección económica?

IX. Números índices

1. Índices simples. 2. Índices de precios agregados no ponderados. 3. Índices de precios agregados ponderados. 4. Índice de cantidades agregadas ponderadas. 5. Índices de valor.

El término *índice* se utiliza en una variedad de contextos. Existen índices que han sido calculados para medir el grado de humedad del medio ambiente, la efectividad del sistema de enseñanza, los cocientes de inteligencia, la actividad bursátil, cambios en los precios al consumidor y muchos otros.

Los números índices son probablemente las medidas estadísticas que más se utilizan; y esto se debe a varias razones. En primer lugar, permiten comparar dos o más series de tiempo que tienen diferentes unidades de medida. Por ejemplo, a través de los números índices podemos comparar los cambios en la producción de automóviles con los respectivos cambios en sus precios. En segundo lugar, mediante los números índices se puede reducir números de magnitud considerable a cantidades manejables, sobre todo cuando se trata de cifras monetarias en situaciones inflacionarias. En tercer lugar, los números índices permiten comparar cambios en la producción de un conjunto de artículos, los que no pueden expresarse en la misma unidad de medida. Por ejemplo, el índice del producto bruto interno pre-

tende medir la variación cuantitativa de todos los bienes producidos en el país. Finalmente, los números índices se utilizan para remover ciertas influencias (*e.g.*, índices de estacionalidad) o para estimar el ingreso real o poder de compra del dinero.

Un número índice es una medida que expresa el nivel de un precio, cantidad o valor en un *período dado* en términos del nivel de la misma variable en un *período base* establecido. Todos los números índices se expresan en porcentajes relativos que tienen el nivel del período base como 100. Supongamos que los precios por galón de la gasolina de 95 octanos en enero, abril y agosto fueron 1.20, 1.80 y 1.90 dólares, respectivamente. Entonces, tomando enero como el período base, tendremos:

$$\frac{\text{Precio de abril}}{\text{Precio de enero}} * 100 = \frac{1.80}{1.20} * 100 = 150$$

El índice de precios de la gasolina de 95 octanos para abril es 150, tomando como período base el mes de enero. De manera similar, calculamos el índice de precios para agosto y nos da 158.

Un número índice para un producto individual, como el índice de precios de la gasolina de 95 octanos, se llama *número índice simple*. Un índice que se calcula para un grupo de productos es un *número índice agregado o compuesto*. Estos últimos son ampliamente utilizados por las empresas como medidas sumarias de costos y precios de venta, de cantidades producidas o vendidas y de valores de producción y ventas.

Los números índices también son producidos y usados por el gobierno y sus diferentes agencias. Por ejemplo, el Instituto Nacional de Estadística e Informática dedica grandes recursos para recolectar información y preparar la medida agregada más conocida de precios, el índice de precios al consumidor (IPC). Otros índices que producen las agencias de gobierno incluyen los índices de precios del PBI, de exportaciones, de importaciones, etcétera.

En este capítulo se presenta la manera de calcular tanto los índices simples como los índices agregados para precios, cantidades y valores. Para ilustrar estos cálculos, a lo largo de todo el capítulo utilizaremos el caso de la empresa comercializadora "Estrella S.A.", que vende tres productos básicos: arroz, azúcar y harina de trigo. El cuadro 28 muestra las ventas y precios de estos productos en los meses de enero, febrero y marzo de este año.

CUADRO 28: PRODUCTOS COMERCIALIZADOS POR "ESTRELLA S.A."

	Precios (Dólares por tonelada)			Cantidades (Toneladas)		
Producto	Enero	Febrero	Marzo	Enero	Febrero	Marzo
Arroz	236.27	240.10	244.90	20,000	22,000	23,000
Azúcar	221.50	230.36	230.36	10,000	9,000	12,000
Harina	352.15	394.41	408.50	5,000	3,000	3,000

1. ÍNDICES SIMPLES

Un índice simple es aquel que compara el precio, cantidad o valor de un producto o artículo individual en un período dado con el precio, cantidad o valor que se registró en el período base.

A. Cálculo de índices simples de precios

La fórmula general para el índice simple de precios o precio relativo está dada por:

$$I_{p,n} = \frac{p_n}{p_0} * 100 \qquad (1)$$

donde p_n es el precio del artículo en el período dado, n.
p_0 es el precio del artículo en el período base.
$I_{p,n}$ es el índice de precios para el período n.

Debemos señalar que el período de análisis puede ser cualquier unidad de tiempo: día, semana, mes o año. Asimismo, notamos que los índices están identificados por suscritos; el primero nos indica el tipo de índice, en este caso p, precio, seguido por un segundo suscrito que define el período para el cual se calcula.

Para el caso de la empresa "Estrella S.A." calculamos números índices simples de precios para cada uno de los productos que comercializa con el fin de analizar el comportamiento de los precios a través de los tres primeros meses del año. Se considerará el mes de enero como período base. Tomando como ejemplo el mes de febrero (f), el índice simple de precios para el arroz será:

$$I_{p,f} = \frac{240.10}{236.27} * 100 = 101.6$$

Los índices para los meses restantes y para cada uno de los productos se calcularon de manera similar, y se presentan en el cuadro 29.

CUADRO 29: PRECIOS E ÍNDICES SIMPLES DE PRECIOS DE LOS PRODUCTOS COMERCIALIZADOS POR "ESTRELLA S.A."

	Arroz		Azúcar		Harina	
Mes	Precio	Índice	Precio	Índice	Precio	Índice
Enero	236.27	100.0	221.50	100.0	352.15	100.0
Febrero	240.10	101.6	230.36	104.0	394.41	112.0
Marzo	244.90	103.7	230.36	104.0	408.50	116.0

La elección del período base es arbitraria. Podríamos haber elegido cualquier otro mes como nuestra base y expresado todos los precios como un porcentaje del precio de ese mes.

Los números índices facilitan la interpretación del movimiento de los precios a través del tiempo. Como se puede observar en el cuadro 29, mientras el precio del arroz subió 3.7%, el precio de la harina se incrementó en 16% entre los meses de enero y marzo.

B. Cálculo de índices simples de cantidad

El índice simple de cantidad o cantidad relativa se define como:

$$I_{q,n} = \frac{q_n}{q_0} * 100 \qquad (2)$$

donde q_n es la cantidad del artículo en el período dado, n.
q_0 es la cantidad del artículo en el período base.
$I_{q,n}$ es el índice de cantidad para el período n.

Los índices simples de cantidad para los productos de la empresa "Estrella S.A." se presentan en el cuadro 30. Como se puede observar en dicho cuadro, las ventas de arroz se incrementaron en 15% de enero a marzo, mientras que las cantidades comercializadas de harina disminuyeron en 40% en el mismo período.

CUADRO 30: CANTIDADES VENDIDAS E ÍNDICES SIMPLES DE CANTIDAD DE LOS PRODUCTOS COMERCIALIZADOS POR "ESTRELLA S.A."

	Arroz		Azúcar		Harina	
Mes	Cantidad	Índice	Cantidad	Índice	Cantidad	Índice
Enero	20,000	100.0	10,000	100.0	5,000	100.0
Febrero	22,000	110.0	9,000	90.0	3,000	60.0
Marzo	23,000	115.0	12,000	120.0	3,000	60.0

C. Cálculo de índices simples de valor

Un índice simple de valor o valor relativo se define como:

$$I_{v,n} = \frac{p_n q_n}{p_0 q_0} * 100 \qquad (3)$$

donde $p_n q_n$ es el valor del artículo en el período dado, n.
$p_0 q_0$ es el valor del artículo en el período base.
$I_{v,n}$ es el índice simple de valor para el período n.

Utilizando los datos del cuadro 28, calculamos los índices de valor (precio * cantidad) para el arroz, azúcar y harina tomando como base los valores de las ventas del mes de enero. Así, el índice simple de valor del mes de febrero (f) para el arroz será:

$$I_{v,f} = \frac{236.27 * 20{,}000}{240.10 * 22{,}000} * 100 = 111.78$$

Este índice indica que el valor de las ventas de arroz en febrero fue 11.78 % mayor que las ventas de enero. El cuadro 31 muestra los índices de valor de los productos comercializados por "Estrella S.A." para los tres meses, tomando como período base el mes de enero.

CUADRO 31: VALOR DE VENTAS E ÍNDICES SIMPLES DE VALOR DE LOS PRODUCTOS COMERCIALIZADOS POR "ESTRELLA S.A."

	Arroz		Azúcar		Harina	
Mes	Valor	Índice	Valor	Índice	Valor	Índice
Enero	4'725,400	100.0	2'215,000	100.0	1'760,750	100.0
Febrero	5'282,200	111.8	2'073,240	93.6	1'183,230	67.2
Marzo	5'632,700	119.2	2'764,320	124.8	1'225,500	69.6

El cuadro 31 permite analizar el comportamiento del valor de las ventas de los productos comercializados por "Estrella S.A.". El valor de las ventas de azúcar se incrementó en 24.8%; las ventas de arroz sólo subieron en 19.2%; y las de harina disminuyeron en 30.4% entre los meses de enero y marzo.

2. ÍNDICES DE PRECIOS AGREGADOS NO PONDERADOS

Ahora consideremos la forma de representar las variaciones de precios de un grupo de artículos. Para obtener un índice de precios agregados no ponderados tenemos dos alternativas. Primero, podemos sumar simplemente los precios del grupo de artículos para el período dado y compararlos con la suma de los precios observados en el período base. Segundo, podemos calcular índices simples de precios para cada uno de los artículos individuales, y luego usar el promedio de estos índices como un índice agregado de precios; es decir, calculamos un promedio de precios relativos.

Supongamos que tenemos una serie de observaciones de los precios de un grupo de k artículos, para un cierto período, y tomamos el período 0 como período base. De acuerdo con la primera alternativa mencionada, el índice de precios agregados no ponderados estará definido por:

$$I_{pa} = \frac{\sum p_{ni}}{\sum p_{oi}} * 100 \qquad (4)$$

donde I_{pa} es el índice de precios no ponderados.

Esta alternativa de cálculo del índice agregado no ponderado tiene una seria limitación. Los precios de los artículos se refieren necesariamente a una unidad de medida específica, y los k artículos podrían estar expresados en una variedad de unidades, haciendo que esta metodología de cálculo sea irrelevante.

La segunda alternativa, el promedio de los precios relativos, está definida por:

$$I_{pr} = \frac{1}{k} \sum \frac{p_{ni}}{p_{oi}} * 100 \qquad (5)$$

donde I_{pr} es el índice de precios no ponderados usando el promedio de precios relativos. El cuadro 32 presenta los cálculos para obtener el promedio de los precios relativos para el caso de la empresa "Estrella S.A.". Los números índices simples de precios para los tres productos se tomaron del cuadro 29.

CUADRO 32: PROMEDIO DE PRECIOS RELATIVOS DE LOS PRODUCTOS COMERCIALIZADOS POR "ESTRELLA S.A."

Mes	Índices simples de precios			Promedio de precios relativos
	Arroz	Azúcar	Harina	
Enero	100.0	100.0	100.0	100.0
Febrero	101.6	104.0	112.0	105.9
Marzo	103.7	104.0	116.0	107.9

Si bien esta manera de calcular los índices de precios agregados ha superado la limitación de la influencia de las unidades de medida del I_{pa}, aún persiste su ineficacia para medir la tendencia general de los precios de los k artículos, pues asigna la misma ponderación a cada uno de estos k artículos sin considerar el volumen de ventas de cada uno de ellos.

3. ÍNDICES DE PRECIOS AGREGADOS PONDERADOS

Para evitar los problemas señalados en la sección anterior, consideremos la posibilidad de usar la información de las cantidades transadas (q) como ponderación en el cálculo de los índices

de precios agregados. Surge el interrogante: ¿estas cantidades se referirán al período base o al período n para el cual se calcula el índice? Dependiendo del período que se seleccione, se definen diferentes índices de precios agregados ponderados, como se verá a continuación.

A. Índice de precios de Laspeyres

El índice de Laspeyres es calculado utilizando las cantidades asociadas con el período base como ponderación de los precios, y está definido por:

$$I_{p,n}(L) = \frac{\sum p_{ni} q_{oi}}{\sum p_{oi} q_{oi}} * 100 \quad (6)$$

Una característica importante en la construcción de este índice es que requiere únicamente información de las cantidades del período base. Esto es muy valioso, en especial cuando la recolección de información referida a cantidades en cada período es imposible o muy costosa. Por otro lado, esto puede ser una desventaja, si es que por alguna razón las cantidades transadas en el período escogido como base no son representativas de la serie de tiempo bajo análisis. Esta dificultad se torna crucial cuando se continúa calculando el índice de Laspeyres para series de tiempo cada vez mayores sin cambiar el período base. La estructura de las cantidades transadas podría cambiar sustancialmente a través del tiempo de tal manera que la estructura original resulte completamente desactualizada. Una manera de resolver este problema es construyendo índices de precios de Laspeyres móviles, en los cuales el período base es cambiado cada cierto tiempo.

Los índices de Laspeyres para el caso de la compañía "Estrella S.A." se calculan tomando como base los datos de precios y cantidades del cuadro 28. Para el mes de febrero (f):

$$\sum p_{fi}q_{oi} = 240.10 * 20{,}000 + 230.36 * 10{,}000 + 394.41 * 5{,}000$$
$$= 9'077{,}650$$

$$\sum p_{oi}q_{oi} = 236.27 * 20{,}000 + 221.50 * 10{,}000 + 352.15 * 5{,}000$$
$$= 8'701{,}150$$

$$I_{p,f}\ (L) = \frac{9'077{,}650}{8'701{,}150} * 100 = 104.3$$

De manera análoga, el índice de precios de Laspeyres para el mes de marzo (m) es:

$$I_{p,m}\ (L) = \frac{9'244{,}100}{8'701{,}150} * 100 = 106.2$$

Esto quiere decir que los precios de los productos comercializados por "Estrella S.A.", ponderados por las cantidades vendidas en enero, se han incrementado en 6.2% entre enero y marzo.

B. Índice de precios de Paasche

En contraste con el índice de Laspeyres, el índice de precios de Paasche utiliza la información de las cantidades de cada período dado como ponderaciones de los precios.

$$I_{p,n}\ (P) = \frac{\sum p_{ni}q_{ni}}{\sum p_{oi}q_{ni}} * 100 \qquad (7)$$

La ventaja de incluir las cantidades transadas en cada período es que ello permite considerar los posibles efectos de sustitución en las cantidades transadas como resultado de cambios en los precios relativos y en los hábitos de consumo. Por otro lado,

como ya se dijo anteriormente, en muchas circunstancias es difícil o muy costoso recolectar información de cantidades para cada período.

La interpretación de los índices de Paasche para cada período es menos clara que aquella de los índices calculados con la fórmula de Laspeyres. En estos últimos, los costos totales que se comparan de período a período son aquellos que corresponden a las cantidades transadas en el período base; en cambio, en el índice de Paasche las cantidades varían de período a período.

Consideremos nuevamente el ejemplo de la empresa "Estrella S.A.". Los índices de precios de Paasche para los meses de febrero (f) y marzo (m) son los siguientes:

$$I_{p,f}(P) = \frac{\sum p_{fi} q_{fi}}{\sum p_{oi} q_{fi}} * 100 = \frac{8'580{,}940}{8'247{,}890} * 100 = 104.0$$

$$I_{p,m}(P) = \frac{\sum p_{mi} q_{mi}}{\sum p_{oi} q_{mi}} * 100 = \frac{9'622{,}520}{9'148{,}660} * 100 = 105.2$$

Los índices de precios de Paasche nos indican que los precios de los productos comercializados por "Estrella S.A." se incrementaron en 4% en el mes de febrero y en 5.2% entre enero y marzo.

4. ÍNDICE DE CANTIDADES AGREGADAS PONDERADAS

Los índices de cantidades agregadas se construyen con el objeto de medir el cambio en las cantidades transadas de un conjunto de artículos a través del tiempo. La metodología para calcular índices de cantidades de Laspeyres y Paasche es similar a la explicada en la sección anterior.

A. Índice de cantidad de Laspeyres

El índice de cantidad de Laspeyres pondera las cantidades individuales por los precios asociados con el período base:

$$I_{q,n}(L) = \frac{\sum q_{ni} p_{oi}}{\sum q_{oi} p_{oi}} * 100 \qquad (8)$$

Es decir, que el índice de cantidades de Laspeyres para el período en curso (n) es el costo total de las cantidades transadas en dicho período evaluadas a precios del período base, expresado como un porcentaje del costo total de las cantidades del período base.

B. Índice de cantidad de Paasche

El índice de cantidad de Paasche utiliza como ponderación los precios del período en curso (n):

$$I_{q,n}(P) = \frac{\sum q_{ni} p_{ni}}{\sum q_{oi} p_{ni}} * 100 \qquad (9)$$

El índice de cantidad de Paasche para el período n expresa el costo total de las cantidades transadas en dicho período como un porcentaje de lo que hubieran costado las cantidades transadas en el período base si se hubieran comprado a los precios del período en curso, n.

5. ÍNDICES DE VALOR

Así como se han obtenido fórmulas para los índices de precios y cantidades, se puede establecer una ecuación para calcular índices de valor. En estos índices el costo total de los bienes transados en un período dado, evaluados a los precios corrien-

tes, es comparado con el costo total de los bienes transados en el período base, evaluados a los precios de dicho período.

$$I_{v,n} = \frac{\sum q_{ni}p_{ni}}{\sum q_{oi}p_{oi}} * 100 \qquad (10)$$

El índice de valor refleja tanto las variaciones en las cantidades como en los precios a través del tiempo. Estrictamente hablando, este es un índice agregado simple, puesto que los valores (q * p) no han sido ponderados.

● *Ejercicios*

1. Para cada una de las siguientes situaciones, explique cómo el decisor puede construir y usar un número índice que lo ayude a tomar la decisión.

a. Un profesional que desea conocer el poder adquisitivo de la tarifa de servicio que cobra hoy en comparación a la que cobraba hace exactamente un año.

b. El Administrador del Club Lawn Tennis desea aumentar las cuotas de los socios de acuerdo con el monto de inflación de la economía.

c. El propietario de un almacén desea conocer el incremento de sus ventas en los últimos cinco años en comparación al incremento de las ventas de una gran cadena competidora.

2. El índice de precios al consumidor (IPC) calculado mensualmente por el Instituto Nacional de Estadística e Informática, como una medida del costo de los bienes y servicios al consumidor, fue de 102.15 para agosto de 1991, usando como período base diciembre de 1990. Suponga que el salario de Carlos Carpio, un ingeniero de "Peru-Gas", aumentó de 840 dólares en diciembre de 1990 a 1,400 dólares en agosto de 1991. ¿Ha aumentado el salario real del ingeniero Carpio? Explique.

3. Una compañía manufacturera que produce tres productos (A, B y C) tiene la siguiente información de cantidades y precios para 1988 y 1991:

	1988		1991
Producto	Precio	Cantidad	Cantidad
A	35	1,200	1,600
B	25	1,000	800
C	30	400	500

a.Calcule los índices simples de cantidad. ¿Cómo evolucionó el volumen de cada producto entre 1988 y 1991?
b.Calcule el índice de cantidades agregado ponderado. ¿Cuál es el significado de este índice de cantidad?

4. Se tiene la siguiente información de los precios del algodón en rama y de los principales insumos requeridos para su producción, para los años 1975 y 1980.

Año	Producto ($/Tm)	Semilla ($/kg)	Nitrógeno ($/kg)	Maquinaria ($/h)	Mano de obra ($/jornal)
1975	18,799	8.54	27.16	188.0	105.40
1980	107,980	56.04	171.78	2,772.68	1,174.49

Calcule los índices simples de precios tanto para el producto de algodón en rama, como para sus principales insumos. ¿Qué ha pasado con los precios en este período?

5. La empresa ENCI tiene la siguiente información respecto a la cantidad y precio de la importación de trigo en el Perú para los años 1975, 1978 y 1980:

Años	Tm	Precio prom. (US$/Tm)
1975	768,100	176.60
1978	789,900	133.90
1980	841,736	194.42

a. Calcule los índices simples de precios y cantidad. Comente estos valores.
b. Calcule el índice de valor para las exportaciones de este producto. Compare estos valores con los de la parte (a).

6. El cuadro que se presenta a continuación indica los precios y las cantidades de tres importantes productos agrícolas exportados por un cierto país de Latinoamérica:

Producto	Precios (Dólares por tonelada) 1989	1990	Cantidad (Toneladas) 1989	1990
Azúcar	103.84	109.98	1'070,315	1'265,874
Café	1,023.15	873.72	28,639	25,182
Cacao	556.75	431.84	34,434	29,115

a. Calcule los índices simples de precios, cantidad y valor de los productos exportados por este país. Explique su significado.
b. Calcule el índice de precios agregado no ponderado.
c. Construya el índice de precios de Laspeyres.
d. Construya el índice de cantidad de Paasche.
e. Construya un índice de valor

7. Una ferretería vende pintura para exteriores y para interiores. Sus ventas de los meses de agosto de 1979 y 1991, y los precios en dólares para agosto de 1979, son los siguientes:

Tipo de pintura	Precio en agosto 1979 (Dólares)	Número de galones vendidos: Agosto 1979	Agosto 1991
Exteriores	11.45	220	290
Interiores	9.67	175	245

Calcule el índice de cantidades agregado ponderado para agosto de 1991, tomando como período base agosto de 1979.

8. La agencia de viajes "Tumi" requería, en 1985, 12 minutos en promedio para hacer una reservación en una aerolínea, 1 hora para

hacer una reservación en un crucero y 18 minutos para hacer otras reservaciones. En 1990, con la introducción de un equipo automático más eficiente, los tiempos para las reservaciones en aerolíneas y otras reservaciones se redujeron a 6 y 12 minutos respectivamente. Sin embargo, el tiempo para hacer una reserva en un crucero no mejoró y continúa siendo una hora. Si en una semana típica de 1990 la agencia de viajes procesa 150 reservaciones de avión, 8 reservaciones en crucero y 37 de las otras reservaciones, ¿cuál es el índice de productividad para 1990, considerando 1985 como el año base?

Apéndice A: Tabla de distribución de Poisson

$$P(x) = \frac{\mu^x e^{-\mu}}{x\,!}$$

Ejemplo: Si $\mu=3$, $X=7$, entonces $P\,(x=7) = 0.0216$

	μ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.00
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659	0.3679
2	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647	0.1839
3	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494	0.0613
4	0.0000	0.0001	0.0002	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111	0.0153
5	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0007	0.0012	0.0020	0.0031
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0005
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

	μ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.3662	0.3614	0.3543	0.3452	0.3347	0.3230	0.3106	0.2975	0.2842	0.2707
2	0.2014	0.2169	0.2303	0.2417	0.2510	0.2584	0.2640	0.2678	0.2700	0.2707
3	0.0738	0.0867	0.0998	0.1128	0.1255	0.1378	0.1496	0.1607	0.1710	0.1804
4	0.0203	0.0260	0.0324	0.0395	0.0471	0.0551	0.0636	0.0723	0.0812	0.0902
5	0.0045	0.0062	0.0084	0.0111	0.0141	0.0176	0.0216	0.0260	0.0309	0.0361
6	0.0008	0.0012	0.0018	0.0026	0.0035	0.0047	0.0061	0.0078	0.0098	0.0120
7	0.0001	0.0002	0.0003	0.0005	0.0008	0.0011	0.0015	0.0020	0.0027	0.0034
8	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0003	0.0005	0.0006	0.0009
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0009

(sigue)

(viene de la página anterior)

x	μ 2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.2572	0.2438	0.2306	0.2177	0.2052	0.1931	0.1815	0.1703	0.1596	0.1494
2	0.2700	0.2681	0.2652	0.2613	0.2565	0.2510	0.2450	0.2384	0.2314	0.2240
3	0.1890	0.1966	0.2033	0.2090	0.2138	0.2176	0.2205	0.2225	0.2237	0.2240
4	0.0992	0.1082	0.1169	0.1254	0.1336	0.1414	0.1488	0.1557	0.1622	0.1680
5	0.0417	0.0476	0.0538	0.0602	0.0668	0.0735	0.0804	0.0872	0.0940	0.1008
6	0.0146	0.0174	0.0206	0.0241	0.0278	0.0319	0.0362	0.0407	0.0455	0.0504
7	0.0044	0.0055	0.0068	0.0083	0.0099	0.0118	0.0139	0.0163	0.0188	0.0216
8	0.0011	0.0015	0.0019	0.0025	0.0031	0.0038	0.0047	0.0057	0.0068	0.0081
9	0.0003	0.0004	0.0005	0.0007	0.0009	0.0011	0.0014	0.0018	0.0022	0.0027
10	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0004	0.0005	0.0006	0.0008
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

x	μ 3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1397	0.1304	0.1217	0.1135	0.1057	0.0984	0.0915	0.0850	0.0789	0.0733
2	0.2165	0.2087	0.2008	0.1929	0.1850	0.1771	0.1692	0.1615	0.1539	0.1465
3	0.2237	0.2226	0.2209	0.2186	0.2158	0.2125	0.2087	0.2046	0.2001	0.1954
4	0.1734	0.1781	0.1823	0.1858	0.1888	0.1912	0.1931	0.1944	0.1951	0.1954
5	0.1075	0.1140	0.1203	0.1264	0.1322	0.1377	0.1429	0.1477	0.1522	0.1563
6	0.0555	0.0608	0.0662	0.0716	0.0771	0.0826	0.0881	0.0936	0.0989	0.1042
7	0.0246	0.0278	0.0312	0.0348	0.0385	0.0425	0.0466	0.0508	0.0551	0.0595
8	0.0095	0.0111	0.0129	0.0148	0.0169	0.0191	0.0215	0.0241	0.0269	0.0298
9	0.0030	0.0040	0.0047	0.0056	0.0066	0.0076	0.0089	0.0102	0.0116	0.0132
10	0.0010	0.0013	0.0016	0.0019	0.0023	0.0028	0.0033	0.0039	0.0045	0.0053
11	0.0003	0.0004	0.0005	0.0006	0.0007	0.0009	0.0011	0.0013	0.0016	0.0019
12	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006
13	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

(sigue)

(viene de la página anterior)

μ

x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0679	0.0630	0.0583	0.0540	0.0500	0.0462	0.0427	0.0395	0.0365	0.0337
2	0.1393	0.1323	0.1254	0.1188	0.1125	0.1003	0.1005	0.0948	0.0894	0.0842
3	0.1904	0.1852	0.1798	0.1743	0.1687	0.1631	0.1574	0.1517	0.1460	0.1404
4	0.1951	0.1944	0.1933	0.1917	0.1898	0.1875	0.1849	0.1820	0.1789	0.1755
5	0.1600	0.1633	0.1662	0.1687	0.1708	0.1725	0.1738	0.1747	0.1753	0.1755
6	0.1093	0.1143	0.1191	0.1237	0.1281	0.1323	0.1362	0.1398	0.1432	0.1462
7	0.0640	0.0686	0.0732	0.0778	0.0824	0.0869	0.0914	0.0959	0.1002	0.1044
8	0.0328	0.0360	0.0393	0.0428	0.0463	0.0500	0.0537	0.0575	0.0614	0.0653
9	0.0150	0.0168	0.0188	0.0209	0.0232	0.0255	0.0280	0.0307	0.0334	0.0363
10	0.0061	0.0071	0.0081	0.0092	0.0104	0.0118	0.0132	0.0147	0.0164	0.0181
11	0.0023	0.0027	0.0032	0.0037	0.0043	0.0049	0.0056	0.0064	0.0073	0.0082
12	0.0008	0.0009	0.0011	0.0014	0.0016	0.0019	0.0022	0.0026	0.0030	0.0034
13	0.0002	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013
14	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005
15	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002

μ

x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0311	0.0287	0.0265	0.0244	0.0225	0.0207	0.0191	0.0176	0.0162	0.0149
2	0.0793	0.0746	0.0701	0.0659	0.0618	0.0580	0.0544	0.0509	0.0477	0.0446
3	0.1348	0.1293	0.1239	0.1185	0.1133	0.1082	0.1033	0.0985	0.0938	0.0892
4	0.1719	0.1681	0.1641	0.1600	0.1558	0.1515	0.1472	0.1428	0.1383	0.1339

(sigue)

(viene de la página anterior)

5	0.1753	0.1748	0.1740	0.1728	0.1714	0.1697	0.1678	0.1656	0.1632	0.1606
6	0.1490	0.1515	0.1537	0.1555	0.1571	0.1584	0.1594	0.1601	0.1605	0.1606
7	0.1086	0.1125	0.1163	0.1200	0.1234	0.1267	0.1298	0.1326	0.1353	0.1377
8	0.0692	0.0731	0.0771	0.0810	0.0849	0.0887	0.0925	0.0962	0.0998	0.1033
9	0.0392	0.0423	0.0454	0.0486	0.0519	0.0552	0.0586	0.0620	0.0654	0.0688
10	0.0200	0.0220	0.0241	0.0262	0.0285	0.0309	0.0334	0.0359	0.0386	0.0413
11	0.0093	0.0104	0.0116	0.0129	0.0143	0.0157	0.0173	0.0190	0.0207	0.0225
12	0.0039	0.0045	0.0051	0.0058	0.0065	0.0073	0.0082	0.0092	0.0102	0.0113
13	0.0015	0.0018	0.0021	0.0024	0.0028	0.0032	0.0036	0.0041	0.0046	0.0052
14	0.0006	0.0007	0.0008	0.0009	0.0011	0.0013	0.0015	0.0017	0.0019	0.0022
15	0.0002	0.0002	0.0003	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009
16	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001

	μ									
x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0137	0.0126	0.0116	0.0106	0.0098	0.0090	0.0082	0.0076	0.0070	0.0064
2	0.0417	0.0390	0.0364	0.0340	0.0318	0.0296	0.0276	0.0258	0.0240	0.0223
3	0.0848	0.0806	0.0765	0.0726	0.0688	0.0652	0.0617	0.0584	0.0552	0.0521
4	0.1294	0.1249	0.1205	0.1162	0.1118	0.1076	0.1034	0.0992	0.0952	0.0912
5	0.1579	0.1549	0.1519	0.1487	0.1454	0.1420	0.1385	0.1349	0.1314	0.1277
6	0.1605	0.1601	0.1595	0.1586	0.1575	0.1562	0.1546	0.1529	0.1511	0.1490
7	0.1399	0.1418	0.1435	0.1450	0.1462	0.1472	0.1480	0.1486	0.1489	0.1490
8	0.1066	0.1099	0.1130	0.1160	0.1188	0.1215	0.1240	0.1263	0.1284	0.1304
9	0.0723	0.0757	0.0791	0.0825	0.0858	0.0891	0.0923	0.0954	0.0985	0.1014
10	0.0441	0.0469	0.0498	0.0528	0.0558	0.0588	0.0618	0.0649	0.0679	0.0710
11	0.0245	0.0265	0.0285	0.0307	0.0330	0.0353	0.0377	0.0401	0.0426	0.0452
12	0.0124	0.0137	0.0150	0.0164	0.0179	0.0194	0.0210	0.0227	0.0245	0.0264
13	0.0058	0.0065	0.0073	0.0081	0.0089	0.0098	0.0108	0.0119	0.0130	0.0142
14	0.0025	0.0029	0.0033	0.0037	0.0041	0.0046	0.0052	0.0058	0.0064	0.0071

(sigue)

(viene de la página anterior)

15	0.0010	0.0012	0.0014	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0033
16	0.0004	0.0005	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0013	0.0014
17	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006
18	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

μ

x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0059	0.0054	0.0049	0.0045	0.0041	0.0038	0.0035	0.0032	0.0029	0.0027
2	0.0208	0.0194	0.0180	0.0167	0.0156	0.0145	0.0134	0.0125	0.0116	0.0107
3	0.0492	0.0464	0.0438	0.0413	0.0389	0.0366	0.0345	0.0324	0.0305	0.0286
4	0.0874	0.0836	0.0799	0.0764	0.0729	0.0696	0.0663	0.0632	0.0602	0.0573
5	0.1241	0.1204	0.1167	0.1130	0.1094	0.1057	0.1021	0.0986	0.0951	0.0916
6	0.1468	0.1445	0.1420	0.1394	0.1367	0.1339	0.1311	0.1282	0.1252	0.1221
7	0.1489	0.1486	0.1481	0.1474	0.1465	0.1454	0.1442	0.1428	0.1413	0.1396
8	0.1321	0.1337	0.1351	0.1363	0.1373	0.1382	0.1388	0.1392	0.1395	0.1396
9	0.1042	0.1070	0.1096	0.1121	0.1144	0.1167	0.1187	0.1207	0.1224	0.1241
10	0.0740	0.0770	0.0800	0.0829	0.0858	0.0887	0.0914	0.0941	0.0967	0.0993
11	0.0478	0.0504	0.0531	0.0558	0.0585	0.0613	0.0640	0.0667	0.0695	0.0722
12	0.0283	0.0303	0.0323	0.0344	0.0366	0.0388	0.0411	0.0434	0.0457	0.0481
13	0.0154	0.0168	0.0181	0.0196	0.0211	0.0227	0.0243	0.0260	0.0278	0.0296
14	0.0078	0.0086	0.0095	0.0104	0.0113	0.0123	0.0134	0.0145	0.0157	0.0169
15	0.0037	0.0041	0.0046	0.0051	0.0057	0.0062	0.0069	0.0075	0.0083	0.0090
16	0.0016	0.0019	0.0021	0.0024	0.0026	0.0030	0.0033	0.0037	0.0041	0.0045
17	0.0007	0.0008	0.0009	0.0010	0.0012	0.0013	0.0015	0.0017	0.0019	0.0021
18	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
19	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0003	0.0003	0.0003	0.0004
20	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001

(sigue)

(viene de la página anterior)

x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
					μ					
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0025	0.0023	0.0021	0.0019	0.0017	0.0016	0.0014	0.0013	0.0012	0.0011
2	0.0100	0.0092	0.0086	0.0079	0.0074	0.0068	0.0063	0.0058	0.0054	0.0050
3	0.0269	0.0252	0.0237	0.0222	0.0208	0.0195	0.0183	0.0171	0.0160	0.0150
4	0.0544	0.0517	0.0491	0.0466	0.0443	0.0420	0.0398	0.0377	0.0357	0.0337
5	0.0882	0.0849	0.0816	0.0784	0.0752	0.0722	0.0692	0.0663	0.0635	0.0607
6	0.1191	0.1160	0.1128	0.1097	0.1066	0.1034	0.1003	0.0972	0.0941	0.0911
7	0.1378	0.1358	0.1338	0.1317	0.1294	0.1271	0.1247	0.1222	0.1197	0.1171
8	0.1395	0.1392	0.1388	0.1382	0.1375	0.1366	0.1356	0.1344	0.1332	0.1318
9	0.1256	0.1269	0.1280	0.1290	0.1299	0.1306	0.1311	0.1315	0.1317	0.1318
10	0.1017	0.1040	0.1063	0.1084	0.1104	0.1123	0.1140	0.1157	0.1172	0.1186
11	0.0749	0.0776	0.0802	0.0828	0.0853	0.0878	0.0902	0.0925	0.0948	0.0970
12	0.0505	0.0530	0.0555	0.0579	0.0604	0.0629	0.0654	0.0679	0.0703	0.0728
13	0.0315	0.0334	0.0354	0.0374	0.0395	0.0416	0.0438	0.0459	0.0481	0.0504
14	0.0182	0.0196	0.0210	0.0225	0.0240	0.0256	0.0272	0.0280	0.0306	0.0324
15	0.0098	0.0107	0.0116	0.0126	0.0136	0.0147	0.0158	0.0169	0.0182	0.0194
16	0.0050	0.0055	0.0060	0.0066	0.0072	0.0079	0.0086	0.0093	0.0101	0.0109
17	0.0024	0.0026	0.0029	0.0033	0.0036	0.0040	0.0044	0.0048	0.0053	0.0058
18	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019	0.0021	0.0024	0.0026	0.0029
19	0.0005	0.0005	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014
20	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004	0.0005	0.0005	0.0006
21	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003
22	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

(sigue)

(viene de la página anterior)

x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.0378
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	1.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.6062	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0034	0.0037	0.0028	0.0031
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

(sigue)

(viene de la página anterior)

x	μ = 11	12	13	14	15	16	17	18	19	20
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0010	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0037	0.0018	0.0008	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.0102	0.0053	0.0027	0.0013	0.0006	0.0003	0.0001	0.0001	0.0000	0.0000
5	0.0224	0.0127	0.0070	0.0037	0.0019	0.0010	0.0005	0.0002	0.0001	0.0001
6	0.0411	0.0255	0.0152	0.0087	0.0048	0.0026	0.0014	0.0007	0.0004	0.0002
7	0.0646	0.0437	0.0281	0.0174	0.0104	0.0060	0.0034	0.0018	0.0010	0.0005
8	0.0888	0.0655	0.0457	0.0304	0.0194	0.0120	0.0072	0.0042	0.0024	0.0013
9	0.1085	0.0874	0.0661	0.0473	0.0324	0.0213	0.0135	0.0083	0.0050	0.0029
10	0.1194	0.1048	0.0859	0.0663	0.0486	0.0341	0.0230	0.0150	0.0095	0.0058
11	0.1194	0.1144	0.1015	0.0844	0.0633	0.0496	0.0355	0.0245	0.0164	0.0106
12	0.1094	0.1144	0.1099	0.0984	0.0820	0.0661	0.0504	0.0368	0.0259	0.0176
13	0.0926	0.1056	0.1099	0.1060	0.0956	0.0814	0.0658	0.0509	0.0378	0.0271
14	0.0728	0.0905	0.1021	0.1060	0.1024	0.0930	0.0800	0.0655	0.0514	0.0387
15	0.0534	0.0724	0.0885	0.0989	0.1024	0.0992	0.0906	0.0786	0.0650	0.0516
16	0.0367	0.0543	0.0719	0.0866	0.0960	0.0992	0.0963	0.0884	0.0772	0.0646
17	0.0237	0.0383	0.0550	0.0713	0.0847	0.0934	0.0963	0.0936	0.0863	0.0760
18	0.0145	0.0256	0.0397	0.0554	0.0706	0.0830	0.0909	0.0936	0.0911	0.0844
19	0.0084	0.0161	0.0272	0.0409	0.0557	0.0699	0.0814	0.0887	0.0911	0.0888
20	0.0046	0.0097	0.0177	0.0286	0.0418	0.0559	0.0692	0.0798	0.0866	0.0888
21	0.0024	0.0055	0.0109	0.0191	0.0299	0.0426	0.0560	0.0684	0.0783	0.0846
22	0.0012	0.0030	0.0065	0.0121	0.0204	0.0310	0.0433	0.0560	0.0676	0.0769
23	0.0006	0.0016	0.0037	0.0074	0.0133	0.0216	0.0320	0.0438	0.0559	0.0669
24	0.0003	0.0008	0.0020	0.0043	0.0083	0.0144	0.0226	0.0328	0.0442	0.0557

(sigue)

(viene de la página anterior)

25	0.0001	0.0004	0.0010	0.0024	0.0050	0.0092	0.0154	0.0237	0.0336	0.0446
26	0.0000	0.0002	0.0005	0.0013	0.0029	0.0057	0.0101	0.0164	0.0246	0.0343
27	0.0000	0.0001	0.0002	0.0007	0.0016	0.0034	0.0063	0.0109	0.0173	0.0254
28	0.0000	0.0000	0.0001	0.0003	0.0009	0.0019	0.0038	0.0070	0.0117	0.0181
29	0.0000	0.0000	0.0001	0.0002	0.0004	0.0011	0.0023	0.0044	0.0077	0.0125
30	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0013	0.0026	0.0049	0.0083
31	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007	0.0015	0.0030	0.0054
32	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0004	0.0009	0.0018	0.0034
33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0010	0.0020
34	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0012
35	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0007
36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004
37	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
39	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001

Apéndice B: Áreas bajo la curva normal estandarizada

Ejemplo:
Si z=1.96, entonces
$p(0 \leq z \leq 1.96) = 0.4750$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767

(sigue)

(viene de la página anterior)

2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Apéndice C: La distribución t de Student (Valores críticos de t)

Ejemplo: Si el área de la cola derecha es 0.05 y los grados de libertad son 10, entonces t = 1.81

Grado de libertad	Área de la cola derecha					
	0.25	0.1	0.05	0.025	0.01	0.005
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95

(sigue)

(viene de la página anterior)

16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79
26	0.68	1.31	1.71	2.06	2.48	2.78
27	0.68	1.31	1.70	2.05	2.47	2.77
28	0.68	1.31	1.70	2.05	2.47	2.76
29	0.68	1.31	1.70	2.05	2.46	2.76
30	0.68	1.31	1.70	2.04	2.46	2.75
40	0.68	1.30	1.68	2.02	2.42	2.70
60	0.68	1.30	1.67	2.00	2.39	2.66
∞	0.67	1.28	1.64	1.96	2.33	2.58

Apéndice D: Tabla de números aleatorios

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09266	64419	29457
10078	28073	85389	50324	14500	15562	64105	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

(sigue)

(viene de la página anterior)

73864	83014	72457	22682	03033	61714	88173	90835	00634	85169
66668	25467	48894	51043	02365	91726	09365	63167	95264	45643
84745	41042	29493	01836	09044	51926	43630	63470	76508	14194
48068	26805	94595	47907	13357	38412	33318	26098	82782	42851
54310	96175	97594	88616	42035	38093	36745	56702	40644	83514
14877	33095	10924	58013	61439	21882	42059	24177	58739	60170
78295	23179	02771	43464	59061	71411	05697	67194	30495	21157
67524	02865	39593	54278	04237	92441	26002	63835	38032	94770
58268	57219	68124	73455	83236	08710	04284	55005	84171	42596
97158	28672	50685	01181	24262	19427	52106	34308	73685	74246
04230	16831	69085	30802	65559	09205	71829	06489	85650	38707
94879	56606	30401	02602	57658	70091	54986	41394	60437	03195
71446	15232	66715	26385	91518	70566	02888	79941	39684	54315
32886	05644	79316	09819	00813	88407	17461	73925	53037	91904
62048	33711	25290	21526	02223	75947	66466	06232	10913	75336

Bibliografía

ÁVILA ACOSTA, Roberto B.
1987 *Estadística elemental.* Eds. RA, Lima.

BARBANCHO, Alfonso
1975 *Estadística elemental moderna.* 4ª edición. Ed. Ariel S.A., Barcelona.

BECHTOLD, Brigitte y Ross JOHNSON
1989 *Statistics for Business and Economics.* PW-Kent Pub. Co., Boston.

CAMPBELL, Stephen K.
1981 *Equívocos y falacias en la interpretación de estadísticas.* Limusa S.A., México

CAO GARCÍA, Ramón
1985 *Métodos estadísticos. Teoría y práctica.* South Western, Cincinnati.

CARLSON, Roger A.
1973 *Statistics*. Holden-Day Inc., San Francisco, Cal.

DAVENPORT Jr., Wilbur B.
1970 *Probability and Random Process*. McGraw Hill, Nueva York.

DOWNIE, Norville M.
1986 *Métodos estadísticos aplicados*. 5ª edición. Harla, México, D.F.

DUHNE REINERT, Carlos
1990 *Técnicas estadísticas y administrativas para el aumento de la calidad y la productividad*. Alfaomega, México.

GONZALES GONZALES, Carlos
1991 *Control de calidad*. McGraw Hill, Bogotá.

GOLDBERGER, Arthur S.
1963 *Econometric Theory*. John Wiley & Sons, Nueva York.

GRANGER, C.W.J.
1980 *Forecasting in Business and Economics*. Academica Press, Nueva York.

GRANT, Eugene L. y Richard S. LEAVENWORTH
1987 *Control estadístico de calidad*. CECSA, México.

HADLEY, George
1981 *Probabilidad y estadística. Una introducción a la teoría de decisiones*. Fondo de Cultura Económica, México.

HOEL, Paul G. y Raymond J. JESSEN
1983 *Estadística básica para negocios y economía*. CECSA, México.

HUNTSBERGER, David y Patrick BULLINGSLEY
1983 *Elementos de estadística inferencial*. CECSA, México.

JOHNSON, Robert R.
1990 *Estadística elemental*. Ed. Iberoamérica, México.

JOHNSTON, J.
1975 *Métodos de econometría*. 3ª edición. Vicens-Vives, Barcelona.

LARSON, Harold
1989 *Introducción a la teoría de probabilidades e inferencia estadística*. Ed. Limusa, México.

LEVIN, Richard
1981 *Estadística para la administración*. Prentice-Hall, México.

MASON, Robert D.
1981 *Estadística comercial y economía*. Ed. El Ateneo, Buenos Aires.

MENDENHALL, William
1990 *Estadística para administradores*. 2ª edición. Ed. Iberoamericana, México.

MILL, R.
1980 *Estadística para economía y administración*. McGraw Hill, Bogotá.

NEWBOLD, Paul
1983 *Statistics for Business and Economics*. Prentice Hall, Nueva York.

NIETO DE ALBA, Ubaldo
1973 *Introducción a la estadística. Concepción bayesiana*. Aguilar S.A., Madrid.

PERSON, Ron
1990 *1-2-3 en el mundo de la estadística. Manual de formatos y aplicaciones prácticas*. Macrobit Ed., México.

RAIFFA, Howard
1978 *Análisis de la decisión empresarial.* Fondo Educativo Interamericano S.A., Bogotá.

RICHARDS, Larry y Jerry LA CAVA
1978 *Estadística en los negocios. ¿Por qué y cuándo?* McGraw Hill, Bogotá.

SALINAS ORTIZ, José
1977 "Ilustración de las técnicas econométricas: Análisis y predicción del sector industrial". Reporte interno. Ministerio de Economía y Finanzas, Lima.
1987 "A Comparison of Macroeconomic Model Structures", en B. Hickman, H. Huntington y J. Sweeney, editores: *Macroeconomic Impacts of Energy Shocks.* North Holland, Holanda.

SELLEKAERTS, Willy (ed.)
s/f *Econometrics and Economic Theory: Essays in Honour of Jan Tinberger.* Londres.

SHAO, Stephen
1972 *Estadística para la ciencia administrativa.* McGraw Hill, México.

SPIEGEL, Murray R.
1981 *Estadística.* 2ª edición. Schaum-McGraw Hill, Madrid.

STEWART, Mark B. y Kenneth WALLIS
1984 *Introducción a la econometría.* Alianza Editorial, Madrid.

THEIL, Henri
1971 *Principles of Econometrics.* John Wiley & Sons, Amsterdam.

WALLIS, Kenneth F.
1973a *Introductory Econometrics.* LSE, Gray Mills Publishing Ltda., Londres.

1973b *Topics in Applied Econometrics.* LSE, Gray Mills Publishing Ltda., Londres.

WONNACOTT, Thomas y R. J. WONNACOTT
1981 *Introducción a la estadística.* Ed. Limusa, México.

ZUWAYLIF, Fadil H.
1977 *Estadística general aplicada.* Fondo Educativo Interamericano S.A., México.

ÍNDICE DE CUADROS

ÍNDICE DE GRÁFICOS

Biblioteca Universitaria

Títulos publicados

- El comportamiento humano en las organizaciones / Javier Flórez García Rada
- Decisiones económicas en la empresa / Folke Kafka
- Deuda externa: Del problema a la posibilidad / Hernán Garrido-Lecca
- Casos de exportación / Óscar Jasaui
- Evaluación estratégica de proyectos de inversión / Folke Kafka
- Introducción a los negocios internacionales / David Mayorga y Patricia Araujo
- Contabilidad, finanzas y economía para pequeñas y medianas empresas / Jorge González Izquierdo y Julián Castañeda Aguilar
- Introducción a la banca / David Ambrosini
- Contabilidad intermedia. Tomo I. Estados financieros y cuentas del activo / Esteban Chong
- Principios de empresas estatales y privatización / Augusto Álvarez Rodrich
- Análisis de decisiones en entornos inciertos, cambiantes y complejos / José Salinas Ortiz

De próxima aparición

- Marketing / Mauricio Lerner y Alberto Arana
- Macroeconomía para la empresa / Folke Kafka
- Técnicas estadísticas de predicción aplicables en el campo empresarial / Jorge Cortez
- Macroeconomía de una economía abierta / María Amparo Cruz-Saco Oyague
- Casos sobre decisiones de marketing en empresas peruanas / Gina Pipoli de Butrón



José Salinas Ortiz (Arequipa, 1948) obtuvo el grado de Ph.D. en Sistemas de Ingeniería Económica por la Universidad de Stanford. Es master en Ciencias Econométricas y Economía Matemática por el London School of Economics, master en Ciencias de Sistemas y Matemáticas por la Washington University e Ingeniero Economista de la Universidad Nacional de Ingeniería.

Autor de varios artículos relacionados con sistemas y análisis de decisiones, del libro *Análisis de decisiones en entornos inciertos, cambiantes y complejos*. Además, es consultor en análisis de decisiones y evaluación de proyectos, y como tal ha trabajado para organismos internacionales —BID, CAF, FAO y AID— así como con diferentes compañías internacionales de consultoría. Actualmente es profesor en la Escuela de Postgrado de la Universidad del Pacífico, en el área de métodos cuantitativos para la toma de decisiones, y representa al Strategic Decision Group (SDG) en Sudamérica.